DOCUMENT RESUME

ED 193 25? TM 800 569

AUTHOR	Burda, Mary Anne, Ed.; Sanders, James R., Ed.
TITLE	Practices & Problems in Competency-Based Measurement.
INSTITUTION	National Council on Measurement in Education, Washington, D.C.
PUB DATE	79
NOTE	153p.
AVAILABLE FROM	National Council on Measurement in Education, 1230 17th St., N.W., Washington, DC 20036 ($6.75).
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Certification; Competence; *Competency Based Education; Criterion Referenced Tests; Cutting Scores; Definitions; Measurement Objectives; *Measurement Techniques; Minimum Competencies; Research Needs; *Standards; Testing Problems; Test Reliability; *Test Validity

ABSTRACT

Three basic areas of definition, measurement, and research, voiced as concerns by members of the National Council of Measurement in Education (NCME) to its Task Force on Measurement Problems in Competency-Based Programs, are used as an organizer for a sponsored symposia and the papers in this monograph. The problem of the definition of competency is addressed in two different ways. Chapter IV presents several different perspectives about how standards of performance may be set, while Chapter II addresses itself solely to the definition of the word "competency." Just as the authors of the seven papers differ in their positions concerning the efficacy of competency-based programs, they also use the word "competence" in different ways, which has real consequences for the way in which research and development efforts will be structured. If the construct definition is selected, then the process of validation of the instrumentation must take into account the framework in which the construct is operationalized, as argued in Chapter V. However, if the definition is considered to be that of setting standards of behavior, then different techniques would be used to validate the cut-off point for an instrument. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

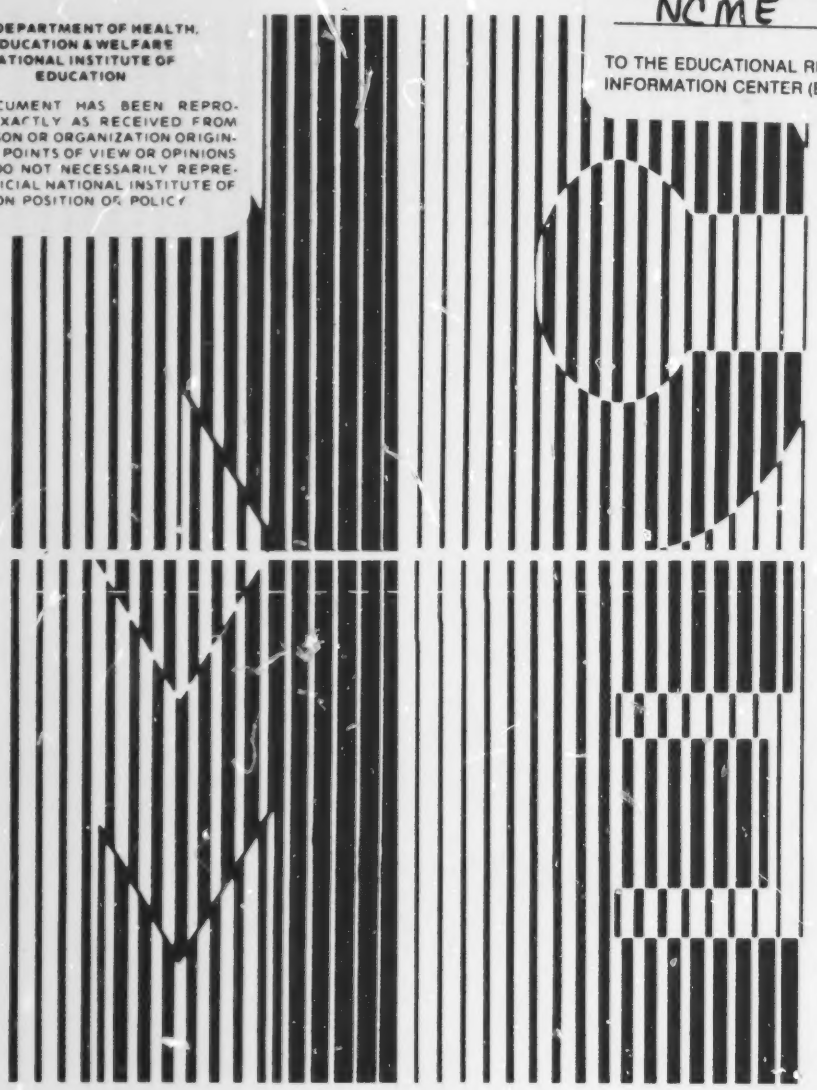
"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
Wm. Russell
NCME
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The National Council on Measurement in Education

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED193255



Practices & Problems in Competency-Based Measurement

Edited by Mary Anne Bunda and James R. Sanders

TM 800569

Practices and Problems

in

Competency-Based Education

edited by

Mary Anne Bunda
and
James R. Sanders

Copyright 1979
by the
National Council on Measurement in Education

All rights reserved

Table of Contents

Acknowledgements v

Chapter I
Introduction 1

Chapter II
A Conception of Competency 10

Chapter III
Practices in Developing Competency Measurement 14

Current Status and Methodological Problems of Validating Professional Licensing and Certification Exams
Kathryn A. Hecht 16

Competence Testing as a Basis for Licensing Problems and Prospects
Paul S. Pottinger 28

Chapter IV
Standards 47

Measurements Consequences of Selected Standard-Setting Models
Richard M. Jaeger 48

Setting Standards
Lorrie A. Shepard 59

Setting Standards in Competency-Based Education: Some Current Practices and Concerns
Larry E. Conaway 72

Chapter V
Instrument Quality Indices in Competency-Based Programs 89

Issues of Reliability in Measurement for Competency-Based Programs
Robert L. Linn 90

Issues of Validity in Measurement for Competency-Based Programs
Robert L. Linn 108

Chapter VI
Messages for Educators 124

References 127

Acknowledgments

This monograph represents the intellectual and spiritual leadership of a vast number of people in NCME. We cannot, as editors, name all of them without the preface of the book being longer than the narrative. There are, however, several individuals who deserve special mention as we trace the history of this document.

The project was initiated under the leadership of Christine McGuire. She and her successors as president of NCME, Tom Fitzgibbon and Joan Bollenbacher, gave support to this effort in terms of funding by the organization and critical review of our plans. Lois Burrill, while she was program chair for NCME, met impossible demands for room and time arrangements. Jason Millman has shown an understanding and patience that we find to be extraordinary. His reminders of our missed deadlines forced the completion of this monograph.

We would also like to extend our thanks to Eva Baker, Jim Impara and Jim Wardrop who served with us on the NCME Task Force. They devoted time to the conceptual efforts of the project and chaired several of the events which we sponsored at the national meetings. In addition, there are about thirty measurement specialists who were involved in programs at the national meeting and hundreds of measurement professionals who responded to our requests in writing who deserve our thanks.

Special thanks to Frank Cookingham who provided assistance to us in our editorial task. Thanks go to Deborah Anzick and Patricia Burness for their fine work in production and design.

Additional Copies can be purchased from

NCME Monograph
1230 17th Street, N.W.
Washington, D.C. 20036

NCME members $ 4.00 plus .75 postage
NCME nonmembers $6.00 plus .75 postage

Chapter I

Introduction

This monograph has been produced under the leadership of the NCME Task Force on Measurement Problems in Competency-Based Programs. To undertand why the editors selected this group of papers, the reader should know the history of the Task Force. During the three years (1974-1977) of its operation, the Task Force interacted with many educators. The dialogue which was created led to a series of issues which the Task Force addressed. This chapter is designed to provide an abbreviated history of the Task Force, a list of concerns which the Task Force compiled, and a description of how this monograph addresses those concerns. It also contains a list of arguments for and against competency-based education. The monograph does not take a position for or against competency-based education. In fact, it contains papers reflective of each side of the issue. The section of this chapter which addresses arguments is designed to present both sides of the issue. Each individual will have to come to a judgment as to which position is the most defensible.

The History of the Task Force

Professional organizations have a special responsibility for the growth of their membership and the solution of problems of special interest to their members. Generally, the vehicle used to accomplish this responsibility is dissemination of current thoughts among the membership via journals and annual meetings. The National Council on Measurement in Education (NCME) took a unique action orientation to its responsibility. Several *ad hoc* groups were formed in the organization and this monograph is the result of one of the groups.

The Beginning

In 1974 Christine McGuire asked, as President of NCME, that a survey of the membership be conducted to identify the areas of interest and concern to the members and also to identify those members of the organization who were willing to work on committees or other *ad hoc* groups in various areas of measurement. The survey resulted in the formation of four Task Forces: Affective Measurement, Teacher Evaluation, Legal Issues, and Problems of Measurement in

Competency-Based Programs. Each of the Task Forces began in much the same way. A chairperson was named for each and in Problems of Measurement in Competency-Based Programs the Chairperson was Mary Anne Bunda of Western Michigan University. The chairperson then invited other individuals to serve on the Task Force. Task Force members were Eva Baker of the Center for the Study of Evaluation at UCLA, James C. Impara, then of the Oregon Department of Education, James R. Sanders, then of the Northwest Regional Educational Laboratory and James L. Wardrop of the University of Illinois at Urbana-Champaign. The Task Force was given the results of the survey and a list of individuals who were likely to be interested in this area of concern. Prior to the 1975 NCME annual meeting, the Task Force surveyed all those individuals who expressed an interest and willingness to work in the area. They were asked to write to Task Force members about specific problems and suggestions and to attend an open hearing at the annual meeting. The program chairperson of NCME made rooms and time available to the Task Force.

At the first open hearing, the audience was mostly made up of individuals in higher education who were working on competency-based programs. However, it became clear from the discussion that competency-based measurement, objective-referenced measurement and criterion-reference measurement were related concerns. People in the audience were concerned not only with the traditional instrument quality indicators, but also were interested in advice on setting the criterion score.

The Adolescence of the Task Force

After the open hearings at the 1975 convention in Washington D.C. were held and the results of a Task Force survey were in, the Task Force had the grist from which it felt it could respond to the membership's demands. It decided to sponsor a series of symposia at the 1976 annual meeting which were issues oriented. As a group the Task Force decided on topics for the symposia and even listed possible paper titles so that the five events would fit together. Each member of the Task Force took responsibility for one symposium. The series of sponsored activities closed with another open hearing. The events which were structured fell into two major categories: A) Issue oriented symposia – Licensing, Measurement Issues, Sociological Issue- and Information Utilization Issues, and B) A debate for which

the resolve was that there is no difference in competency-based program measurement and any other measurement. The events were well attended, but the composition of the audience was changed from the previous year.

Historical events were forcing other people to worry the question of graduation by test. The open hearing left us with three nagging questions—What instruments are being used? How should standards be set? and What are the quality indicators for instruments being used? The plan prior to the convention was to collect the papers from the symposia for a monograph. And, indeed, four papers in this monograph had their start at the 1976 annual meeting. However, it became clear after the presentations and discussions that the Task Force didn't feel ready to put itself out of business by printing a document. It had not brought issues to closure, in fact, it had started several new discussions.

Consequently, the Task Force decided to let the discussion continue among our paper presenters and the audience through the year and sponsored two more events at the 1977 convention. In its naive way, the Task Force still was looking for consensus positions. Of the two sponsored symposia, the first addressed the issues of standard setting and reporting information. The second was concerned with methodologies of measurement other than the traditional paper and pencil technique. After the two symposia were held another Task Force meeting was called. It was clear that consensus did not come and would not come through symposia. The Task Force concluded that the range of positions which were presented needed to be presented to the professional education community.

And So the Work of the Task Force is Over

The papers in Chapters III and IV of this monograph represent a sample of papers which were prepared for the sponsored symposia. The editors chose to highlight in this monograph the three major issues which evolved during the course of the Task Force life, i.e., practices in measurement in competency-based situations, standard setting and technical issues. The editors could not reprint all of the papers which were given as part of the symposia due to volume. Exclusion of any paper should not be viewed as a negative judgment of the paper, rather it is due to limitations of space. The Linn papers in Chapter V of the monograph are a combination of work which he

did for the Task Force and other papers written under government funding while he was a visiting scholar at the Center for the Study of Evaluation at UCLA. Chapter II of the monograph is designed to give the reader an understanding of the various positions that were taken on the measurement of competence. The definition of "competence" changes in this monograph and Chapter II is designed to alert the reader to the various ways it can be used.

Why has it Taken So Long?

Through the life of the Task Force, the Task Force members felt that somehow they would be able to represent a consensus point of view on the methodological problems and other issues of competency-based programs. Task Force members saw the results of the Task Force being a "how to" book of readings which operationalized that consensus point of view. They waited. Consensus did not arrive in the measurement community and other political and educational groups kept expanding the scope of the problem. Consequently, the editors decided to publish a wide range of perspectives before the papers became untimely. Our belief is that this collection represents the best thoughts of the field on the practices and problems of competency measurement.

Competency-Based Programs

While this monograph is designed to address the issue of measurement problems in competency-based programs, it is difficult to discuss the value of a measure without discussing the value of the program. The monograph does not take a stand for or against competency-based programs. However, individual authors reflect both points of view. Rather than have the reader encounter these points of view in a context other than a discussion of the program itself, the arguments — both pro and con — are drawn together here. The order of the presentation and the number of points listed in the argument should not be taken as bias. We hope that proponents of each side will see a just representation of their position. The reader will have to decide which position is most defensible in his or her own context.

Arguments for Competency-Based Programs

1. Specification of exact skills that students should possess upon exiting school will force teachers and other school personnel to be accountable.

2. Exiting students will have skills which are essential to survival in the world of work and be more capable of securing a job. Consequently, the level of national unemployment will be reduced.

3. The public is demanding usable skills as exemplified in the recent court cases in which a high school graduate has sued a school district because the graduate had a diploma and yet was viewed as incompetent by employers.

4. Programs which have a clear focus upon the outcomes of the school will also clearly focus the use of resources and put a lid on the increasing costs of education.

5. A listing of the skills which a student possesses will provide more information as part of the record than mere grades.

6. Attendance in school is not equivalent to acquisition of skill. There is a need to have students demonstrate competence.

7. Programs can be used as vehicles to respond to the community which the school serves, thus improving the relation between the school and its various publics. Members of the community and potential employers of the students can become involved in the determination of the competencies.

Arguments Against Competency-Based Programs

1. Neither schools nor teachers can control all potentially potent variables affecting the "competence" of students.

2. Such programs will narrow the emphasis of schools on vocational training or simple literacy rather than broad educational goals.

3. Students who previously would have dropped out of school are currently encouraged to stay in. An emphasis on certification or selection may renew the trend toward dropping students from school without graduating them.

4. Emphasis on skill acquisition will negatively affect the socialization function of the school. This movement is counter to the temporal tandem movement to "mainstream."

5. The program will tend to increase the cost of education by requiring schools to remediate with special programs. If students are retained in school enrollments will rise, thus increasing the tax based budgets of schools.

6. If extra resources are not given to the school, revised programs which support the competency instruction will drain resources from special programs, e.g., the gifted students or vocational training. Thus, a social cost will be incurred.

7. There is little research to help in the selection of "essential behaviors," and performance in such desirable areas of life as success or happiness. Substituting achievement of competencies for completion of knowledge based course work would track students early based on what could be irrelevant criteria.

8. Problems of differential performance on tests across races will occur. Note that Title VII of the Federal Civil Rights Act of 1964 prohibits employers from engaging in discriminatory practices in their hiring policies. Under the guidelines, discriminatory use of tests including specific educational requirements have been addressed. If a disproportionate number of minority group applicants fail, employers have the burden of proving job-relatedness of the requirements using validation techniques prescribed by the Equal Employment Opportunity Commission. If a disproportionate number of minority students fail a school testing program, a similar requirement of proof may be imposed on the schools.

9. The record-keeping requirements of the program will be costly in terms of staff time, while the testing requirement of such a program will be costly in terms of student and staff time.

10. School board members, legislators, employers and other key citizens ought not dictate what or how the schools should teach. This is the responsibility of practitioners in education.

11. If current tests and grading procedures do not indicate competence, there is no reason to believe a new test (and grading procedure implied by setting performance standards) will indicate competence.

Issues Identified by the NCME Membership

Review of the preceding lists reveals that the arguments used by both the proponents and opponents of competency-based programs come from a wide variety of perspectives. The complexity of the arguments was also revealed in the survey conducted by the Task Force. The request of the Task Force to interested members was to identify one (or two) of the most important issues in competency measurement. About 40 members of NCME responded to the call for issues. Their responses fell into three general conceptual areas which the Task Force labelled Definition, Measurement and Research. Abstracts of those concerns are listed below.

Conceptual Concerns

Definition area:

1. How to define competency vs. non-competency

2. How to identify characteristics of a trait that are amenable to competency-based measurement systems

3. Need to develop learning theory support for competency-based education

4. Need to develop standards (How competent is competent?)

Measurement area:

1. How to make scores on two tests (which measure the same competency) comparable

2. How to develop instruments for measuring competencies

3. How to develop goal-referenced item/scale pools

4. Need to study interpretability of criterion-referenced data as a function of the mode and method of reporting results, the normative prompts provided, and the length and complexity of the discussion provided

5. Need to develop procedures for determining the validity and reliability of criterion-referenced objectives

6. Need to consider multiple matrix sampling techniques to handle "mini-tests" which are a part of competency-based education

7. Need to consider new statistical tools, since the usual central tendency and dispersion tools may be inappropriate for criterion-referenced data

8. Need to determine what comprises sufficient measurement for individual decision-making

9. How to use competency measurement for placement testing (or diagnosis)

10. What alternative measurement techniques can be used in competency testing

11. Teaching teachers what we know about measuring competencies

12. Areas of measurement of competence needs to be expanded beyond the cognitive domain

13. How to generalize across time and settings (problem of transfer)

14. How to measure higher level cognitive competence

15. How to deal with the logistics of individual competency testing

Research area:

1. Impact of teacher on pupil learning

2. Prediction of success based on acquisition of assumedly needed competencies

3. Need to survey and hear from those who have labored through competency-based programs

4. How to validate the competencies themselves

Responses to Concerns

The three basic areas of concern were used as an organizer for the sponsored symposia and also as an organizer for the papers in this monograph. As the Task Force read the requests in the Research

area, they looked to areas in which testing had a history of being used to select individuals and certify them. Thus, the Hecht and Pottinger papers in this monograph are presented so that school personnel may use the experience of those who are involved with licensing. The very specific requests in the Measurement area will not be addressed. However, Chapter V of this monograph does contain a comprehensive review of procedures which may be used to judge the technical quality of measures used in competency-based programs and many of the other papers refer to techniques used in the construction of measures.

The problem of the definition of competency will be addressed in two different ways. Chapter IV presents several different perspectives about how standards of performance may be set, while Chapter II addresses itself solely to the definition of the word "competency". Just as the authors of the papers differ in their positions concerning the efficacy of competency-based programs, they also use the word "competence" in different ways. In fact, the varying definition of the word leads to problems of communication in a discussion of the issues.

Chapter II

A Conception of Competency

Law-makers, school board members, and other public representatives who affect the operation of public education in this country are moving toward and accepting competency-based education as a comprehensive, contemporary reform of education. The Education Commission of the States has reported that by March, 1978, thirty-three states had established some form of competency requirement in public elementary and secondary education. Considering the growing acceptance of this reform it behooves professional educators to develop and communicate a clear and understandable meaning of the term "competency." If competency is to be the primary goal of elementary and secondary education, it is inconceivable that professional educators can continue to practice, not knowing exactly what "competency" means.

Many might argue that concern over the meaning and operationalization of competency-based education is outside the domain of expertise of measurement specialists. This is a curriculum concern, a policy concern, a political issue, the argument might go. However, in response to such reasoning, the measurement specialist must note that if competency is to be a crucial outcome of education, one must be able to determine whether it exists. And, in order to capture it, one must know what he/she is looking for. There is no way to develop competency measures without first arriving at an adequate definition of the concept. Adequacy in this instance refers to a definition so clear that everyone would agree that they recognized it if it indeed existed. Arriving at an adequate definition of competency is a tall order considering the lack of careful examination of the concept apparent in the semilities surrounding public debate and discussion of competency-based education. A review of the elements of competency requirements across states and school districts suggests that the rule for defining competency is that anyone can define it in any way they please as long as they state what they mean. And, in many cases, the qualifier (". . .as long as. . .") is missing. Based on the experience of the NCME Task Force, it is safe to conclude that there is currently no prevailing technical definition of "competency."

The problem that is encountered when searching for a definition of competency is that there are too many. The definition depends on the source. Below are several definitions of the term from our correspondence that we have sorted into two types:

Competence (Type A):

1. A student's progress depends on competency.

2. Competencies mean knowledge, skills and behaviors that a student must have.

3. Competencies are ability, skill and fitness.

4. Competencies are abilities.

5. An overall characterization of the individual is needed rather than the knowledge of possession of discrete competencies.

6. Competencies are either physical, personal (i.e., personality) or mental. Physical is the result of natural endowments enhanced by practice; personal the result of experience, initiation, and adaptive behavior modifications; cognitive the result of assimilation of useful information to form a structure of knowledge and understanding.

7. Skills, behaviors, and information which are directly related to specific functions that students will perform as workers or citizens after graduation.

8. Competencies are learner behavior objectives.

9. Competencies are institutional goals.

10. Competencies are based on what people do.

Competence (Type B):

1. We need a way to separate the competent from the incompetent.

2. Criteria for determining competence are explicit and public.

3. Performance is the major source of evidence of competence.

4. Rate of progress through a program is determined by demonstrating competence (not determined by time, semester hours or

some other similar standard).

5. Competence is required demonstration of mastery.

6. Competency is the adequacy for the task.

7. Competency-based programs are the result of the realization of publicly specified criteria for classes of learning outcomes found to be appropriate to student preparation, i.e., knowledge, skills (performance) and products.

8. Demonstrated ability to perform to criterion and function at a specified job level.

9. A label given to the results of a comparison of a particular performance standard or behavioral criterion. Competencies are inferences, then, and each competency is only inferentially related to the learning process from which the static comparison is derived.

10. An interaction where it approximates a standard or the behavioral criterion.

11. Minimum skills in personal development, social responsibility, career development.

While there is quite a bit of variance within each of the categories, there is a common way of viewing the term "competence" in each group. The **Type A** definitions conceive of competence as a hypothetical construct. Much like the words "skill,""achievement," and "intelligence" constructs, "competency" in this use fits into some conceptual framework. When curriculum specialists talk of "collecting lists of competencies," they are using the term to refer to a construct. However, the breadth of the construct definition varies greatly. In some uses of the construct definition, competency is broader than the word "skill" and refers to a combination of cognitive, affective, and psychomotor skills. Other individuals use competency as synonomous with behavioral objective which is generally a restrictive definition of a skill.

The **Type B** definitions of competence refer to a **standard of performance** either implicitly or explicitly. In this use, the term closely parallels definitions of mastery or criterion levels of performance. Educators have asked, "At what level of performance does a student

become competent?" or "How do we distinguish between competent and incompetent performance?"

The choice of one definition as opposed to the other has real consequences for the way in which the measurement community will structure research and development efforts in competency-based measurement. If the construct definition is selected, then the process of validation of the instrumentation must take into account the framework in which the construct is operationalized. The arguments for the selection of this definition are eloquently given in Linn's validity paper in Chapter V of this monograph. However, if the definition is considered to be that of setting standards of behavior, then different techniques would be used to validate the cut-off point for an instrument.

Chapter III

Practices in Developing Competency Measures

One way of gaining insight into the implications of adopting a competency-based program in education would be to investigate the experiences of those who have been working on the measurement of competencies. Although there is little to look at in the area of elementary and secondary education, there has been a relatively great amount of attention given to the measurement of competencies in professions such as nursing, law and medicine, and in many vocational areas. One reason for the great amount of activity in these areas has been the movement toward licensing and certification of members of the professions and trades. Those working in elementary and secondary education might profit from looking at their collective experience and by asking themselves, "What can be learned from a review of areas with a relatively long history of research and development in competency measurement, and what does this history say as we consider adopting competency-based programs?"

The first paper in this section provides a glimpse of competency test use for licensing purposes in certain professions and trades. Dr. Hecht has reviewed test development practices for a variety of licensing exams with a concentration on validation procedures. Some interesting background information about the advent and use of these exams leads one to ponder whether a similar context now surrounds the growing interest in competency measures in elementary and secondary education.

The second article deals with procedures for defining the competencies to be measured (using the construct definition of competence) and alternative types of measures for assessing proficiency as a competency. Dr. Pottinger makes numerous references to work on competency measurement in education and reasons for attention being directed to the measurement of competencies. It is interesting to note the recency of most references, giving us a good indication that we are only beginning to develop a technology of competency measurement. When one finds a technical procedure developed in 1971 referred to as "classical," he/she can take that as a sign of technological immaturity.

The two papers in this section should have a sobering influence on those who would want to use competency measures in elementary and secondary education.

Current Status and Methodological Problems of Validating Professional Licensing and Certification Exams

Kathryn A. Hecht
Center for Cross Cultural Studies
University of Alaska, Fairbanks

In 1973, I became involved in the licensure validation issue when I was asked by the National League for Nursing to prepare a background paper on the validation of the RN (registered nurse) licensing examination and related work on performance testing. Naively, I thought it would be a simple task of consolidating what had been done in other professions. That was a far more complex and interesting task than I had expected, and questions and concerns raised during the assignment led me to realize the need for further interaction among measurement specialists on this topic. As far as I know, the 1974 symposium, entitled "Validation of Professional Licensing and Certification Examination: A Methodological Dilemma," was the first such discussion sponsored by NCME. I felt at that time that the symposium was quite successful. I am pleased to see those issues discussed again, two years after the first symposium, and consider it to be an opportunity for follow-up, although the context seems a bit different.

Let me briefly review some of the points discussed in my earlier paper. First of all, I felt it was important to define the difference between licensure and certification as I use those terms. I prefer Jensen's (1972) distinction where he discusses licensure and certification as two types of minimum competency testing in that the purpose of the tests is to establish an individual's status in relation to an established go/no-go criterion. Licensing is usually a mandatory program designed to protect the public from incompetent practitioners; that is, to prevent an individual with particular deficiencies from entering practice. Jensen calls this a "selecting-out" process. Certification, on the other hand, is usually a voluntary program where the emphasis is on granting special status to an individual with more than run-of-the-mill knowledge, ability and/or skill. Jensen calls this "selecting-in." Perhaps the best known example of a selecting-out exam would be for a driving license, where the public is protected from those whose knowledge is judged to be below adequate

standards. An example of selecting-in, or certification, would be the "diplomate" program for medical specialities or recent certification programs for automobile mechanics. Since validation deals with the purpose for which the test is intended, I believe these distinctions to be important. Unfortunately, the words are often used interchangeably. For example, I believe teaching certification to be a misnomer, according to this definition, because it is a legal requirement to begin teaching, to protect the public from incompetent teachers, and signifies no special standing within the profession.

Next, I took a brief look at the sudden growing interest about validation of licensing and identified four possible reasons for this concern. First was the criticism of testing in general, which in the past decade has become a popular, headline-making cause. Second, there was a proliferation of jobs requiring licensing, and the hodgepodge of state and local legislative bodies emerging to control the process. (Shimberg, Esser and Kruger's 1972 report entitled *Occupational Licensing and Public Policy* is the only comprehensive document I have seen on the subject and it provides an excellent overview). Third, the civil rights movement continued to make inroads against discrimination, especially those efforts focused on discriminatory hiring practices. Fourth, challenges were being made to many professions to obtain status through alternative training routes rather than the traditional curriculum or school-based methods. This implied a question of who qualifies to take a licensing exam. In 1972, for instance, many returning army medics sought to take the RN exam and were denied access on the grounds of not having graduated from nursing school. I think it is still safe to say that licensing is experiencing a period of attention and questioning for a number of reasons, not the least of which is federal legality. Licensing agencies continue to be subject to challenge to prove their tests are valid predictors of job performance, measuring job-related skills.

Despite growing concern about the topic, I found an incredible lack of information, particularly research information. This was especially true in trying to relate licensure to job performance. The information I was able to locate was scanty, often considered in progress, and being done in subject matter areas rather than considered collectively as a methodological problem. Very seldom was material available through professional journals and, in some cases, professions considered such information confidential.

Something that Maslow said in advising the Council on Occupational Licensing hit home:

> I am convinced that we need to sharpen our ability to develop and demonstrate the rational relationship between the job requirements and the measurement system used to certify or qualify people for an occupation. A number of techniques are available to improve the process of job analysis to get a much more exact fix on the critical requirements for the work to be done. I would urge, therefore, that especially in examinations for occupational knowledge and proficiency, you insist, at the very least, on a clearcut showing of how one proceeds from the decision as to the skills and abilities required for effective performance to the decisions that certain or other measures will insure that the applicant can adequately perform in that occupation. (1971, p.339)

Recent Events Related to Validating Licensure Exams

What changes there have been on some of these points over the past few years! First of all, there has been some professional attention directed at this problem which I had found lacking previously. Let me cite the APA publication *Principles for the Validation and Use of Personnel Selection Procedures* (1975) as an example. Issued by the Division of Industrial-Organizational Psychology of APA, it is intended to accompany the APA publication *Standards for Educational and Psychological Tests* (1974). However, while this publication says it intends to provide:

1. principles upon which personnel research may be based;

2. guidance for practitioners conducting validation; and

3. Information which may be interpreted for personnel managers and others who may be responsible for authorizing or implementing validation efforts.

It also says that the principles are not intended to be a technical translation of existing or anticipated legislation. I note this publication as evidence of heightened professional interest, but still there seems to be a reluctance to relate methodological principles, professional outlook, etc. to the predicaments in which many professionals now find themselves in relation to validating licensing and certification exams. We still seem to have methodologists in one corner and applied researchers or practitioners in the other. Practitioners are

working in subject fields, while most methodologists are dealing with general principles. I have found little evidence that the "twain has met" in the past few years.

Concerning the availability of information, it appears to me that there is almost as little overall information available now as there was a few years ago on how one should go about validating a licensing or certification exam, and on methodological issues related to these procedures. There has been little in the professional journals dealing with this issue as a practical problem or even methodologically as a general problem.

Another thing that does not seem to have changed is the trend toward the proliferation of licensing exams. For instance, the *Readers Guide* (1973-74) lists under "licensure" articles with the following titles: "Should Auto Mechanics be Licensed By Law? Pro and Con Discussion;" "Licensing Proposal for Parenthood;" "Debate Over Licensing for Consultants;" "How to Strike Back Against Appliance Repair Abuse – Suggestion for Mandatory Licensing;" "Drivers Licenses for Snowmobiles?" The problem of building and validating licensing tests doesn't seem to be going away – it seems to be increasing!

Within the professions, there appears to be evidence of continued interest in certification exams – exams that grant recognition to a specialty area. For example, the American Nurses Association is now in the process of developing a series of certification exams in nursing specialties, such as one already available in Psychiatric-Mental Health nursing (1974). In 1971, the National Association of Social Workers introduced a written objective examination for certification in their field. I am sure there are numerous others.

In terms of legal challenges, I do not feel adequately versed to deal with this issue. However, the work of the Equal Employment Opportunity Commission, the Civil Service Commission and others involved is very important to our topic, and needs to be carefully observed and interpreted. Unfortunately, information on what is happening with regard to the legal situation – and especially how that might relate to methodologists – is also scarce.

The Civil Rights Digest devoted its Spring 1975 issue to Job Discrimination and Affirmative Action. White, in an article, "Testing and Equal Opportunity," noted: "In its recent decision in *Moody v. Albemarle Paper Co.*, the Supreme Court kept up the momentum of equal employment opportunity by favoring compensatory back pay awards and by specifying appropriate standards of proof in determining whether tests are job related" (1975, p.49).

The author was actually referring to a 1973 circuit court opinion on *Moody v. Albemarle* which said, with respect to the testing question:

> We think Albemarle has failed in several respects to show that its tests are job related, have a manifest relationship to employment, and have been validated in accordance with EEOC guidelines. . .
>
> In developing criteria of job performance by which to ascertain the validity of its tests, Albemarle failed to engage in any job analysis. Instead, test results were compared with possible subjective ratings of supervisors who were given a vague standard by which to judge job performance. Other courts have expressed skepticism about the value of such ill-defined supervisor appraisals. (1973, p.139)

However, in June, 1975, the Supreme Court did uphold and expand upon the circuit court ruling that Albemarle had not proved the job relatedness of its testing program (*Albemarle v. Moody*, 1975). There have been additional significant court decisions since 1975; for an update see Ledvinka and Schoenfeldt (1978). The Albemarle decision gave great weight to the EEOC *Guidelines* (U.S. Government, 1974) and also referenced the APA *Standards* (1974). The Court quoted the *Guidelines* concerning use of rankings:

> The Guidelines provide (that) the work behaviors or other criteria of employee adequacy which the test is intended to predict or identify must be fully described; and, additionally, in the case of rating techniques. . .whatever criteria are used they must represent major or critical work behavior as revealed by careful job analysis (1975, p. 25).

Problems in Validation of Licensure Exams

Let me now address the problem as I see it and as viewed over the past few years. One part of the problem is that validation studies of licensure exams are rare indeed. Seldom is the test development process that sophisticated or comprehensive. Often what is offered is some sort of content validation methodology resulting from such diverse methods as (1) calling in a group of professional cohorts to review questions to (2) doing fairly adequate analysis of critical skills. Seldom are such exercises reported except to say they did exist.

It would appear to me that predictive criterion-related validation studies would be the type most closely fitting the expressed purpose of licensure exams: that of assuring minimal competency on the job for the protection of the public. Interest is with the criterion not yet obtainable at the time of testing -- to predict an individual's outcome prior to a situation occurring when that person is on the job. Of course, it can be said that for licensure exams, such validation studies, the predictive type, are difficult to develop, time consuming, impractical for numerous reasons, and expensive. It can also be said that psychometric methodology offers little guidance for such studies; the area of licensure lacks classic studies familiar to those who were schooled in psychological testing. Once all this is comprehended, the fact that such validations are rare, almost non-existent, is not surprising but nevertheless disconcerting.

Research has shown that course grades are not efficient predictors of job performance (best demonstrated by the Navy in World War II: Gulliksen, 1950; Stuit, 1947). We know, for instance, that many written tests are found to correlate only too well with reading tests in non-professional skill areas. We also know there is a lack of relationship between achievement as measured by paper and pencil tests and performance measures. This has been demonstrated in such diverse professions as education (Quirk, Witten, and Weinberg, 1972), and engineering (Hemphill, 1963). This lack of correlation suggests great importance for validation of licensure tests. As far back as 1951, Ryans and Frederiksen clearly summed this point up from the measurement perspective:

From the standpoint of validity one of the most serious errors committed in the field of human measurement has been that which assumes the high correlation of knowledge of facts and principles on the one hand and performance on the other. Nevertheless, examinations for admission to the bar, for medical practice, for teaching . . . are predominantly verbal tests of fact and principle in the respective fields. Relatively little attention has been paid to the testing of performance as such. (1951, p. 455)

If training and knowledge variables are not necessarily sufficient to predict job proficiency, it appears logical to look next to what is broadly called performance testing. If one accepts the definition of performance testing as being a test which is "relatively realistic" (Fitzpatrick & Morrison, 1971), then this is a logical place to look for the answer to: (1) how to validate licensure exams more effectively, and (2) how to revise licensure exams themselves if necessary.

The most interesting and well-documented use I found concerning performance measures in predictive research was in the area of employee selection and promotion. Of course, the employer situation has numerous advantages over licensure boards, such as: control over subjects, limited range of jobs, job descriptive information, length of period of observation, and the possibility of gradually implementing a testing program, allowing research time to study predictions before changing the decision process.

Assessment centers appear to be an effective performance-based type of employment or promotion screening device. The procedures simulate situations that would be faced if the examinee were to be moved up to a higher grade position, and using a sophisticated form of role playing, develops information regarding how he might cope with decision-making. Validation studies done on assessment center techniques have shown them to be a better indicator of future success than any other tool management has yet devised. (Bray and Campbell, 1968, described how one such center works and included the validation process.) Although the assessment center concept could be used as a validation tool for licensing exams, as an ongoing technique, where large numbers of people must be tested each year, it appears to be unrealistic.

Illustrative of a more practical approach to introducing performance into testing situations are two types of programmed testing developed in the medical profession. In the assessment of clinical competence, performance is simulated on objectively scored paper

and pencil tests. The National Board of Medical Examiners first introduced the concept (Hubbard, 1964) and now use programmed testing for the medical licensing exam Part III on clinical competence, which previously was a practical bed-side type of oral examination. There are two competing models in the medical profession for this type of testing. In both models the examinee is confronted by a realistic clinical situation and proceeds through a series of decision choices, each step accompanied by an increment of information upon which the next depends, similar to programmed instruction. The linear model used by the National Board allows for only one correct option at each decision point while the branching model promoted by McGuire and Babbot (1967) allows for more than one route to a solution and is used in specialty certification.

To my knowledge neither variation of programmed testing has been validated in relation to predicting job performance. The Part III, or clinical competency exam, is said to derive its validity by measuring something different from what Parts I and II measure, which is strongly related to medical school course work; scores on Part III correlate only moderately with those on Part I and II. Cronbach (1970, p. 444), having reviewed this so-called validity evidence on Part III, noted: "Follow-up studies are needed to make sure the test measures a skill of medical practice and not just ingenuity in test taking."

Similar to problems confronted by those attempting predictive validation of licensure tests, performance tests development logically begins with an identification of specific skills and abilities involved in the activities the test is designed to predict. The next step is the choice of representative tasks, a difficult task which strongly influences the validity of the performance tests. Other difficulties with performance testing come from a lack of applied methodology in that performance tests are frequently criterion-referenced and procedures for estimating reliability and validity still are subject to debate.

As noted previously, licensure and certification exams are types of minimal competency exams and, like the performance measures we have been discussing, would normally be considered criterion-referenced. The examinee is often theoretically tested in terms of absolute criteria (how well the person can do what he or she needs to do to perform adequately in the job). Comparison among test takers is usually not the purpose of licensing exams. However, most licensing

exams on which I have reviewed research data still are developed using norm-referenced models.

Most of the literature discusses performance tests as a new form of assessment, used to increase the realism of the test. My primary interest in performance tests is less commonly discussed, that of providing criteria for predictive validation. The only suggestion of such research use I have located was a theoretical discussion, "Providing A Criterion Measure," in the article by Ryans and Frederiksen:

> Performance test data may provide, first of all, a criterion for research. Information yielded by performance tests makes possible the validation of other measures which, although of a more indirect nature, may be more economical in administration. In many situations it is difficult and expensive to administer performance tests to large numbers of examinees. Such situations demand the construction of psychometric instruments that will yield measurements related to criterion and will also be practicable. In the construction of aptitude tests for various skills and operations, performance tests may provide the criterion against which the available second order test can be judged. (1951, p. 466)

Recent Validation Efforts

The National Board of Medical Examiners appears to have no new validation studies available on the National Board Examination Parts I, II and III for medical students. (The two studies sent to me were both over ten years old.) However, in 1974 the National Board did form a R & D Department, and their correspondence indicates several developmental studies which include validation components are in progress.

Two validation studies of certifying exams in the medical profession on which information was supplied concerned practitioners of internal medicine and physician assistants. Schumacher (1973) reported a validation study of the written examination for internal medicine which was based upon comparing test performance of first and third year residents. According to the author, the study attempted to "determine whether the examination is sensitive to gains in knowledge, ability and skill that presumably occur during graduate training in internal medicine" (p. 132). The study was said to "support the hypothesis that the examination had validity for measuring achievement at the graduate level" (p. 133). The study makes no pretense of relating achievement to expected future performance or even a sound job analysis.

The Physician Assistant certifying exam, new in 1973, also was said to have validity in that those with formal training scored higher than those just beginning training. However, that report (National Board of Medical Examiners, 1973-74) also noted evidence of construct validity in that post-graduates with clinical experience after graduation did better on the exam than those without, although experience prior to entering training had not proven significant. An additional study to estimate concurrent validity compared program faculty ratings of students on 40 statements of clinical competence with examination performance. Four of the six examination components correlated significantly (though the highest was only .29) with at least one of three rating factors. The report indicated that modifications will be made in the following year to refine the rating scale. However, the report does not deal with predictive validity nor is there any mention of content or job analysis in developing the rating form.

The National League for Nursing[1] is proceeding with plans to validate the RN licensure exam, and in designing the project, has accepted "as a first principle, [that] the sampling of content for a new test should be altered in such a way that items drawn from on-the-job experience would receive precedence over items based mainly on knowledge." The rationale given is that licensure is meant to assure safe and effective practice, and to evaluate nursing practice would require judgments as to whether the decisions and actions taken in sample cases are appropriate. In shifting from curriculum-based learning to job-related behavior, the question becomes how to sample items from the job situation. The League, with assistance from American Institutes for Research, currently is carrying out a critical incident study which will become the empirical base for a criterion-referenced test. Almost 14,000 incidents were collected in five specialty areas with geographic and institutional diversity. Classification of incidents and test specification are the activities of the current phase. The League plans to construct a performance simulation test and compare results with the current RN exam. It currently is seeking funding to develop more elaborate performance simulation exams and to attempt validation by observation of performance. This appears to be the most active professional group on the validation scene to date!

[1] This information is drawn from a draft proposal which the author has permission to use but which is not generally available. (See Hecht, 1974, for additional information on the RN licensure exam and related studies.)

Another center of activity, Educational Testing Service, has been involved in certification or licensure test development for social workers, lawyers and nursing specialists, and their efforts are worth watching. In 1969 the National Association of Social Workers added an objective written test to the membership requirements for the Academy of Certified Social Workers. Since 1973, ETS has been administering that test. Research on the test reported by Boyd (1975) noted that no item clusters were large enough or differentiated enough to provide sub-scores – subjects doing well in one subject area would do well in another. Biographical data also showed Blacks with consistently higher failure rates. Upon further study, Sharon (1974), recommended reduction of jargon and wordiness, that study guides be provided and that the test be empirically validated. A validation study for the social worker certification exam has just been approved. The information I received stated that the exam should grant certification to competent workers, regardless of test-taking ability, and deny certification to incompetent workers regardless of their theoretical knowledge. (Sounds more like a delayed licensure exam!) The validation method proposed uses supervisor rankings on overall competence and on several specific dimensions of job performance related to a revised version of the exam. If pairs of raters agreed, the criterion is assumed to be reliable and items thus developed to relate to effective job performance. No mention of job analysis was made. The plan also stated, among use of study results, that such information will provide evidence regarding fairness in case of litigation.

A validity study on the multi-state bar exam was near completion in 1975 (draft, Carlson & Werts, 1975). The purpose of the study was to investigate the relationship among the Multi-state Bar, bar exam essays, law school grades, law school admission tests and undergraduate scores. The summary noted that reasonable relationships were found among these measures. Although this study described the relationships of the exam to previous achievement/aptitude measures of law students, it did not appear to be concerned with predicting adequate job performance. Therefore, the exam's validity as a licensing instrument appears in need of further study.

No research information was available on the Psychiatric-Mental Health Nursing Certification Examination sponsored by the American Nurses Association and developed by ETS.

Given the research information cited in this and my previous paper, it would appear that most professions have a long way to go in adequately validating licensure and certification exams. Few studies are directed at predictive validity or the performance which can be expected on the job. Many seem content to relate exams to academic measures based on curriculum content and training which have no proven relationship with the job tasks to be performed. Most exams are put into use before validity studies are started or with none in mind until problems identify themselves. And those professions I have cited here are probably further along than most. But given recent legal decisions, time seems to be running out.

An additional challenge upon the horizon to professional individual licensure is institutional licensure. Interestingly, it appears most vocal in the one profession which has appeared to take leadership in licensure validation – nursing, and other health care occupations. Agree (1973) reviewed the movement in health care and noted support from the Department of Health, Education and Welfare in testing various plans of institutional licensure for health personnel. Both the AMA and the NLN have formally opposed institutional licensure in lieu of individual licensure for doctors and nurses. If nothing else, this alternative should further prompt action toward upgrading licensing and ceritfication programs and can be viewed as an added critical pressure on professions to police themselves wisely and equitably.

In closing, I think the field of licensing and certification continues to present methodologists with real and immediate challenges. Here are practical problems based on real and current concerns. If each occupation continues struggling on its own, without serious attempts from groups such as NCME to provide integrated conceptual and methodological frameworks, solutions will remain a long way off. I have seen little to indicate cohesive methodological efforts being made to tackle any of the problems mentioned herein. Professional licensing and certification exams continue to be developed and used without adequate validation data. The work that is being done tends to be within professional fields rather than in methodology, and few share their experiences or the knowledge gained. What is needed is a methodological attack on the underlying problems that impede solutions to dilemmas faced by those responsible for licensure and certifications.

Competence Testing as a Basis for Licensing: Problems and Prospects

Paul S. Pottinger
Institute for Competence Assessment

The common sense notion that an academic credential represents at least a minimal level of competence is hard to overcome. This is true in spite of significant empirical evidence that credentials are not causally linked and often not correlated with actual performance in the world of work (Heath, 1976; Hecht, 1976; Hogan, 1977; Hoyt, 1965; McClelland, 1973; White, 1976).

The negative social impact of licensing based upon credentials (and therefore upon tests which only correlate with academic achievement) has been documented elsewhere (Huff, 1976; White, 1976).

If credentials are not reasonable indicators of non-academic performance, then for purposes of licensing we must look to other indicators of minimal levels of professional performance necessary to protect the public.

The major alternative has been tests. If testing is to substitute for credentials, test scores must be more indicative (i.e., predictive) of professional competence than credentials. This is a tall order. The state of the art of testing is not very advanced with respect to prediction. Thus, it must be emphasized from the start that testing is only a potentially useful strategy. Much research needs to be done.

This paper will review some important aspects of test development that reveal the complexity and tenuousness of testing as a viable alternative for determining who should be allowed to practice in the numerous professions which require licensing.

Other alternatives for licensing (e.g., training) and non-licensing strategies (e.g., registration) should be carefully examined as mechanisms for protecting the public.

In educational and work settings, numerous changes in assessment procedures have occurred in the past ten years as a result of demands for accountability from legislators, the courts and the public. These

changes are reflected in the shift from assessing what people know to what they can do. In the world of work, licensing practices have not kept pace with these changes because the courts, licensing boards or agencies and recipients of licenses have not demanded it. This is beginning to change, however, and the competency-based testing movement should provide licensing agencies and boards with concepts and techniques for making licenses more useful, valid and fair. Assuming for the moment that professions are genuinely interested in protecting the public through licensure (a questionable assumption), the problem of what constitutes protection must be clearly addressed. Let us assume that we can come to some decision about this as a matter of judgment. Once proper consensus is reached about what constitutes protection, it becomes an empirical matter to determine how this is best achieved. One critical aspect of determining what works best will be the proper identification of the knowledge, skills, abilities and other characteristics that define minimal levels of competence necessary to protect the consumer. After these "competencies" have been identified, they must be empirically linked to outcomes. Finally, a means of measuring these qualities reliably, validly and meaningfully is required.

The remainder of this paper will explore issues related to:

techniques and considerations in identifying what constitutes competent performance;

some critical concepts related to measurement techniques for translating these competences into measurable variables;

emprical linkages between assessment techniques and job requirements; and

the current state of the art of licensing exams and what mights be done to improve it.

Techniques and Considerations in Identifying What Constitutes Competence

The determination of what constitutes competence is "the rock upon which all defensible examination procedures must be built" (Loveland, 1976). Test development logically begins with an identification of specific knowledge, skills, abilities and other characteristics which comprise occupational performance domains. The definition of performance domains is a difficult task, but it influences the validity

of tests more than any psychometric consideration in test development. It is remarkable that this is generally accepted, yet so little attention is given to competency identification methodology. The truth is that few licensing test developers have seriously questioned the methods of using subjective expert opinions as the "rock" upon which tests are developed. Rather, we tend to be more concerned about reliability coefficients and content validity, cut-off scores, magnitudes of classification errors, decision rules and parametric assumptions than about the meaning of all these with respect to competence. The result of not understanding in the first place what really constitutes competence has resulted in millions of dollars and person-years of research directed at predicting outcomes of dubious consequence. In trying to establish order and control over our subject matter through the limited technology that characterizes the state of the art, we tend to oversimplify. For some, this means to manipulate test items or perform statistical analyses without understanding the nature of human behavior. For others, it means to simplify complex human behavior into superficial, observable and manageable variables. Only a few have ventured into the world of work with a vision of human performance that is commensurate with its true richness and compexity and who are willing to define competence without resorting to the state-of-the-art techniques which have done so poorly for so long.

What are these techniques for defining competence that represent the state of the art and what are some alternatives?

Expert Consensus

The most popular and dangerous technique for defining competence is to accept the judgment of experts. This statement will arouse strong feelings among many professionals, but the empirical evidence is overwhelming that the phenomena of selective perception, beliefs and value systems contaminate objectivity so as to make expert judgments unacceptable. Even large groups of judges can be wrong; there is no protection against subjectivity in numbers! However, many social scientists and others outside of the social science community who understand the vagaries of subjectivity act as if they do not.

Job Analyses

There are numerous job analysis techniques that have been used to identify the information, skills and other characteristics necessary for competent professional practice. But from the start, one must differentiate these techniques according to two separate but important di-

mensions. These differentiations are critical to predicting who will be competent performers.

1. One must differentiate techniques which identify critical **dimensions of the job** from those which identify critical **characteristics of job performers.**

2. One must differentiate techniques that identify critical job or performer characteristics which are task, situation or **level specific** from those which identify critical job or performer characteristics that are **broad or generalizable** across job and situations and throughout a wide range of career performance levels.

The traditional technique for identifying common or unique elements of competent performance is to perform one of a variety of types of job function analyses. A classical approach developed by Fine and Wiley(1971) is to classify jobs according to continuous job requirements. This job function analysis approach is often based primarily on motor skills analysis and has utility in their identification, but it is too narrow an approach to be used as a method for determining significantly more complex dimensions of professional practice. Variations of this method exist, many of which fall under the rubric of the behaviorial objectives approach. This approach, sometimes carried to extreme, resuits in taxonomies of hundreds, sometimes thousands, of skills connected with particular kinds of jobs. These taxonomies neglect many significant areas of job competence by virtue of the fact that they address only external observable behaviors without consideration for intrapersonal and contextual (environmental) variables with which behaviors interact.

While the job function, task analysis and behavioral objectives approaches may help one understand common job elements for setting "equitable" pay scales, creating some training programs or enabling lateral job mobility, they typically do not differentiate which aspects of the job are most important to competent performance, and, more importantly, they do not identify critical and differentiating characteristics of the job performer. Furthermore, even rigorous behavioral observations fail to adequately assess either complex interactions among variables or underlying causes of competent professional practices.

Many test developers and researchers have been led away from accounting for the richness and complexity of human performance by assumption of behaviorism or by the "tyranny of reliability" (Pot-

tinger, 1976). They have fallen prey to the specific behavioral objectives approach because they believe that molecularizing global behaviors into subunits is theoretically sound. Or they believe that this reductionism makes behavior more amenable to objective observation and definition, which, *ipso facto*, solves the problem of reliability. Implicit (and sometimes explicit) in this approach is the assumption that the acquisition of micro-skills and abilities adds up in some linear and simplisitic fashion to overall competence. That is, highly valued generalizable skills are assumed to be mere summations of the ability to perform objectively measurable and highly specific subtasks.

This preoccupation with specificity, clarity and precision of behavioral objectives has left many researchers with an operational paradigm for defining and measuring professional competencies which is intuitively and theoretically oversimplified and invalid. The resulting lengthy checklists of highly specific "low inference" behavioral outcomes – even when properly sampled from the content/behavior domain, item analyzed and content validated – have been found empirically (though not yet legally) insufficient for defining or validly predicting competence.

Any judgment-based approach may yield reliably observable outcomes, but may provide no insight into the skills and abilities that **cause** these outcomes. Too often qualities of the mind and character not amenable to direct observational techniques account for the presence or absence of competent performance.

Behavioral Event Analysis

A technique for identifying such characteristics of job performers that are causally related to complex criterion outcomes has been developed by Professor David C. McClelland of Harvard University and his colleagues. The behavioral event analysis technique consists of a structured interview. It involves obtaining a number of descriptions of "behavioral episodes." For example, a person might be asked to think of incidents or events pertaining to job performance in which he felt he was particularly effective and then describe in detail what led up to the incident, when and where it occurred, and how he was feeling, thinking and reacting before, during and after it. He would also be asked to describe incidents in which he felt he was ineffective or in which things did not work out the way he hoped they would. Generally each person interviewed would be

asked to report on several effective and ineffective incidents, events or episodes.

A distinguishing characteristic of this interview procedure is that it elicits information from which actual overt and covert behaviors can be reconstructed, rather than eliciting interpretations or perceptually biased recollections of past behavior. Responses are recorded and analyzed by professionals trained in concept formation techniques to yield constructs from the interview data which account for how more effective or less effective practitioners perform their work differently. It is the "latent" content which emerges from these interview protocols, rather than the manifest content, which is important to this concept formation process. Thus, what is perhaps most important about this technique is that it assumes that more important than either the storage and retrieval of information or the frequency and range of tasks performed (as specified in behavioral objectives) are hosts of covert behaviors which are antecedent or are causally related to competent professional behaviors. That is, analysis of the latent content of data leads to definitions of competence in terms of variables that cause behavior rather than in terms of external behaviors that are observable manifestations or "building blocks" of competence. These latter behaviors may correlate with success, but correlation or mere relatedness does not imply causation, a fact almost completely overlooked in the task analysis or behavioral objectives model of competence. Specific behaviors may be statistically associated with professional competence, but it does not follow that these behaviors lead to competence; they may have followed it. This model is aimed toward defining and measuring characteristics which are not only statistically related to outcomes, but in fact cause them.

A concrete example of this model comes from the work of McClelland (1974) and his colleagues with the U.S. Information Agency. It was universally agreed that superior U.S. Information Officers possessed a high degree of communication skills in that they were able to effectively deal with people from different nationalities and backgrounds. Communication skill *per se* is a criterion that could be easily rated with a high degree of reliability. However, it was found that the reason these superior officers could communicate with people so well was that they possessed two other characteristics which permitted them to do so. One was an ability; the other was an

attitude. They had the ability to empathize with people, i.e., to use nonverbal cues as information and to ask questions designed to elicit the real needs of their clientele. In addition, they had a strong positive attitude toward people in general, consisting of the conviction that people are basically good and that they have the capacity to change for the better when given the means to do so. Thus, if the determination of competence were only based upon the demonstration of communication behaviors, it would ignore the critical causal elements that are necessary for superior communication performance as a U.S. Information Officer. Empathy and positive bias are very difficult to observe on the job and therefore communication skills would be the desired observable criterion performance in this example. However, understanding how one achieves this criterion must often take into account attributes or characteristics which are unobservable from the point of view of an observer or even of the person engaged in the task of communicating.

Klemp (1977) recently summarized work that he and his colleagues have done over the last five years involving people in careers as diverse as human service workers, military officers, alcohol abuse counselors, small business entrepreneurs, police officers, process consultants, salesmen, State Department officers, civil service technical and administrative workers and industry executives. He discovered three broad dimensions of performance which were highly related to competence but are rarely considered in licensing exams: cognitive process abilities, interpersonal skills and motivation.

The first category will be discussed later. Interpersonal skills included such characteristics as empathy, diagnostic listening, faith in the ability of people to change and other personal attributes that, for example, allowed one to respond appropriately to clients' needs or to promote feelings of efficacy in another person.

Klemp's review of analyses of competent performers in a variety of occupations provides convincing empirical support to the notion that cognitive, interpersonal and motivational factors are critical to effective performance. What is more important perhaps, is the fact that all aspects of the three dimensions of behavior and personality which he reviews can be identified, defined and measured. Thus, these dimensions should be treated as critical domains of any profession when tests of competence are being developed.

This is particularly true in light of Fortune's (1976) findings that there are three principal causes of failure by professionals to provide competent services: failure to stay current in their field, failure to make personal day-to-day adjustments to practice, and failure in one or more fundamental skill areas needed for competent practice. Some consider the third cause of "malpractice" to be minimized by licensing examination. The other two causes of failure are typically not considered amenable to detection or prevention by examination. Klemp's data indicates that this need not be the case if the domains of performance are adequately defined and properly assessed. For example, the willingness and ability to learn new things, which might be related to staying current in the field, and motivational characteristics, which differentially affect one's ability to adapt to certain kinds of occupational requirements or demand characteristics, can be measured. This fact places these critical aspects of competence within the grasp of license exam developers.

It can be seen from this brief overview of several approaches to defining competence that the domain of behaviors considered important to competent professional practice can be viewed quite narrowly (e.g., in terms of knowledge) or quite broadly (e.g., in terms of complex interactions among knowledge, behaviors and covert variables). It is commonly accepted, however, that any professional practice domain is complex. It is multi-dimensional, i.e., including several dimensions of behaviors such as thoughts, feelings and actions.

The domain is dynamic. Knowledge, skills, abilities and other characteristics that constitute competence at one level of a profession may be quite different from those that constitute competence at another level. That is, what makes a junior physician, teacher or lawyer effective may change as one achieves more senior status. And other changes in the professional domain can be expected as a result of scientific and technical developments or the changing needs of society (Loveland, 1976).

We often hear of the limitations of licensing in accounting for the constantly changing nature of professional job requirements. Continuing education and re-licensing are sometimes proposed as logical solutions to having licensing exams conform to the dynamic nature of job requirements. Still another dynamic aspect of professional practice which receives little attention has to do with the differences among the variety of settings in which practitioners perform.

Research in recent years has demonstrated that environmental variables which variously characterize occupational settings are powerful mediators of job performance. These variables have been repeatedly shown to be closely and causally related to factors such as satisfaction, efficiency and performance (Campbell, Dunnette, Lawler & Weick, 1970; Franklin, 1973; Likert & Bowers, 1969, 1973).

Occupational settings provide different working climates, and people with given sets of skills and abilities will perform very differently in response to these environments. Furthermore, characteristics of an occupational setting place demands upon a practitioner or worker which are independent of one's skills, personal characteristics and the requirements of the job. Thus, even if we could measure knowledge, skills and other attributes of people to appropriate levels of criteria out of the context of the occupational setting, these indicators of competence would not mean that a person could use these competencies effectively within the demands or constraints of the setting in which one performs.

The dynamic nature of the context within which personal characteristics and job performance requirements interact preclude the prediction of even minimally competent performance in all but very narrowly specified interactions of people, job requirements and environmental variables. Such interactions can sometimes be specified for purposes of job selection and training, but a profession or occupation taken in its entirety is not amenable to such specificity. The notion that a profession or occupation is homogeneous is, of course, simple-minded, yet most licensing exams are given under such a tacit assumption. Even the best (i.e., highly predictive) test is only applicable to relatively static environmental conditions. The dynamic nature of occupations, due to change in technical knowledge and social needs and especially due to the heterogeneity of environmental setting variables within professions, greatly limits the utility of tests in assuring minimal competence even on a re-licensing basis.

Finally, any professional competence domain is individualistic. Different practitioners may competently manage the same problem in different ways. On the face of it, this statement seems intuitively obvious, yet our licensing examination procedures rarely allow for any individual differences in reaching common solutions to problems.

Clearly, how one defines the domain of competence will greatly affect one's choice of measurement procedures. In simplest terms, if one considers the domain of competent practice to be comprised exclusively of knowledge, then knowledge tests may be appropriate. Nobody would consider knowledge a sufficient domain of professional practice. Yet the majority of licensing exams are limited to this definition. If one considers behaviors a part of the domain, then tests must evoke behaviors which reflect this aspect of competence. Simulation techniques, for example, attempt to evoke such behaviors. If one considers personal attributes such as empathy, maturity, well-developed cognitive processes, the ability to think clearly under stress, moral reasoning abilities, interpersonal effectiveness or motivation as important aspects of competence, then these domains must also be reflected in the assessment process.

These multi-dimensional, dynamic and individualistic aspects of competence create a plethora of test development problems which affect the format and psychometric properties of the tests which are used.

Critical Concepts Related to Measurement Techniques

Let us examine five measurement concepts that should be considered in developing a test once the performance domain has been properly defined.

Measuring Cognitive Processing Skills Rather Than Knowledge
Psychologists have often failed to develop measuring instruments that are sensitive enough to detect effects of primary interest to professional practitioners and their clients – the consumer. Test developers have unnecessarily restricted the range of methods they have employed to measure knowledge, skills, abilities and other characteristics that comprise competence. Time-saving and money-saving incentives have resulted in a predominance of measures which utilize the multiple-choice questionnaire format.

It would serve us well to ask the extent to which these assessment techniques have any bearing on what people are required to do in their professions and on the competencies people possess. In our daily lives we are constantly called upon to process various kinds of information, to analyze its components, to associate this new information with that which we have stored away in our memories, to

partial out the crucial information from the trivial and to integrate this information into our cognitive structure. In this way we constantly use information from many sources to solve problems, and in the process, we learn new things about our world and ourselves.

In truth, people are almost never asked to recognize a correct answer among a list of three or four alternatives. Rather than being reactive to such well defined situations, people must be proactive in situations which provide only partial information.

The one thing that most traditional testing methods have in common, regardless of what they purport to assess, is this: They only measure one's ability to retrieve information after it has been stored. Many such methods fail even in this. A multiple-choice test, for example, measures the ability to recognize rather than recall.

Storage and retrieval of information should not be of primary importance in licensing. Rather, the issue is a more substantive one: How is the knowledge used to come to grips with the practical problems of performing one's job competently?

Perhaps the most consistent finding that Klemp (1977) discovered in his analysis of professional competence is that the amount of knowledge of a content area is generally unrelated to superior performance in an occupation. Of greater significance, he often found it unrelated even to marginally acceptable performance. The implications of this finding for the majority of current licensing practices is quite obvious. Existing requirements of knowledge are not necessarily required for competence in many occupations. While some knowledge is no doubt required for the satisfactory discharge of work-related duties, Klemp found that more important was that an individual be willing and able to learn new things. Klemp articulated three cognitive skills related to competent performance in a broad array of occupations. One is the ability to see thematic consistencies in diverse information and the ability to organize and communicate those consistencies. A second related cognitive skill is the ability to conceptualize the many sides of a controversial issue. A third cognitive skill that Klemp finds often taken for granted is the ability to learn from experience. While Klemp's work draws from a wide range of occupations, McGuire's (1963) research on cognition in medical practice also reflects the complexity and multi-dimensionality of cognition in

professional practice. She defines seven levels of cognitive domain: recall of isolated information, recognition of meaning or implication of performance, simple interpretation of data or application of a single principle or standard combination of principles, analysis of data or application of a unique combination of principles to a novel problem, evaluation of a total situation, and analysis of a variety of elements of knowledge and application to a novel problem situation in its entirety (Loveland, 1976).

Both Klemp and McGuire have developed and/or identified new measures of these complex aspects of cognition which appear to be related to professional competence in a variety of occupations. Their works provide ample evidence that the majority of current licensing exams do not reflect what is known about competence and can be measured.

The Problem of Method Variance

One reason tests have been so highly criticized is that only a small part of the richness of thinking and behavior is tapped by paper and pencil tests.

There are many personal qualities comprising professional competence that we would like to measure such as common sense, managerial skills, leadership abilities, interpersonal effectiveness, moral reasoning, empathy, initiative, problem-solving abilities and others. Unfortunately, those who award credentials and licenses have often settled for measuring small components of these qualities in terms of specific knowledge, skills and abilities that they hope are related to these more general qualities. One reason for this reduction in measurement is that the technology of ability measurement is not good enough to get at the larger, more consequential characteristics of people.

We can easily fall prey, however, to further reductions in the quality of assessment by limiting ourselves to only one method of measurement. Campbell and Fisk (1959) have documented the common notion that the more one increases different perspectives and techniques in measuring a phenomenon, the better will be the measurement. Typically, in testing for licenses, we have limited ourselves to a set of respondent-type measures. These measures usually require multiple, forced-choices among a set of prepared alternatives in a paper and pencil format. By limiting ourselves to these paper and

pencil tests, we are measuring the effect of the test format as much as we are measuring the knowledge, skills and abilities being assessed. In technical terms, this is the issue of "method variance," i.e., how much we are measuring the method relative to how much we are measuring some personal attribute.

Assessing professional competence by using a series of paper and pencil tests is analogous, for example, to measuring how fast someone can drink by requiring one to use a straw. In this example, the paper and pencil test and the straw are equivalent in that they both limit the phenomenon being measured in a reliable way. We would get a better understanding of true professional ability (as well as the ability to drink quickly) if we worked toward eliminating the constraints of measurement. One way of doing this is to use measures that break away from simple modes of measurement. In doing so, in any case, we must require that the measurement techniques we use are objective and quantifiable.

The Problem of Complex Skill Interactions

Earlier it was noted that people can solve identical problems in different ways, yet tests allow for little if any flexibility in this regard. A related phenomenon to this notion that multiple cognitive, affective and behavioral strategies can lead to the same outcome is the fact that competent skills in one area can sometimes compensate for deficiencies in others. In other words, a variety of combinations of individual performance levels could theoretically add up to equivalent overall performance. Thus, minimal levels of performance on individual variables (which comprise overall competence) may have little meaning in themselves. Their interactions with respect to outcomes may have far greater significance. A common practice in scoring licensing exams, however, is to add up the number of correct items and use the resulting total score as a measure of competence. This procedure is valid only if the examination is unidimensional. A proper examination, however, is multi-dimensional and requires sophisticated techniques for determining the optimal patterns of interactions among subscores.

Some researchers such as Hoffman (1976) and McClelland and Burnham (1976) have been investigating better ways of establishing minimal levels of competence that recognize the importance of interactions. By developing statistical decision rules based on empirically derived algorithms, Hoffman has developed a technique for de-

termining the "optimal syntheses concerning subunits of the performance domain." How much this technique allows for variations in patterns is unclear, but recognizing the importance of patterning as opposed to mere summation is a step in the right direction.

McClelland and Burnham reported the importance of the interaction between levels of motivation and ego-maturity for managerial competence. These efforts and findings reflect the fact that competence is not a simple summation of discretely defined subunits of knowledge and skills or personal characteristics. Most licensing exams, however, are based on such a simple and erroneous assumption about the nature of competence.

The Problem of Maximal Levels

A variation of this problem of oversimplifying the meaning of test scores involves the assumption that if minimal levels of competence can be determined, demonstrations of greater than minimal competence on any one examination subunit is obviously desirable. Licenses are often restricted to those whose test scores are higher than minimal levels required for competent performance. This is especially true when cut-off scores are determined by (1) manpower supply and demand in the professions, (2) the desire to minimize false-positive measurement errors, (3) the desire to "upgrade" the profession, or (4) other "arbitrary" decisions about who should be allowed to enter the professions.

Such occurrences discriminate unfairly against those who are competent but are selected out of occupational opportunities by those who believe in the simple equation: higher test scores mean better job performance. The tacit assumption that superior abilities in all measured skills or characteristics are desirable for performance is highly questionable. While it makes sense to require minimal levels of proficiency for some competencies, ability levels over and above necessary cut-off points do not always correlate with overall performance.

For example, McClelland and Dailey (1974) found that a minimum level of organizational or clerical ability was necessary for competent performance of human service workers in the Massachusetts Civil Service system, but high scores on measures of these abilities were negatively correlated with superior job performance. Similarly, while one can assume that a minimal level of verbal ability is necessary for competent performance as a college teacher, a recent study at Har-

vard University revealed that past SAT scores of faculty members were negatively correlated with more successful teachers. These findings suggest that requiring more than sufficient (minimal level) scores on a license exam to gain entry into the profession could be dysfunctional to "upgrading" the profession as well as unfair.

Generic Skill Vs. Observable Performance Skills

A fifth concept has to do with measuring abilities that are causally related to successful performance rather than being merely correlated with performance. This point was elaborated upon earlier. Suffice it to repeat here that many assessment techniques are based on external behaviors which, although they are the building blocks of successful performances, tend to be reductionistic and lack meaning because they fail to assess underlying causes of these behaviors. This problem has important implications for licensing: The observable behaviors which are assessed in even the most sophisticated licensing examinations fail to adequately predict performance because they do not take into account variables which determine the effective use of these behaviors in practice.

Job Relatedness: Empirical Linkages Between Assessment Techniques and Job Requirements

Let us look more closely at linking assessment techniques with job requirements for competent performance. For convenience of discussion and analysis, one can describe techniques according to a continuum of most to least directly performance related. At one end of the continuum are work sampling measures which consist of transferring on-the-job behaviors directly into the assessment situation. At the other end are measures which can be demonstrated to be statistically related to work performance, although the reasons for this relationship (correlation) need not be clear.

Work Sampling Measures

Two examples in this category involve simulation techniques. For example, with regard to complex sales, managerial and leadership roles, the assessment center approach is a popular example of this type of measurement technique (Bray, 1964; Bray & Campbell, 1968; Bray & Grant, 1966; Byham, 1970; Dunnette, 1971). The assessment center has enormous face validity. Situations that simulate those faced by managers on a day-to-day basis are presented to management candidates. These participants cope with the demands,

pressures and problems of management while being observed and rated by a panel of judges. Participants who react effectively to these simulated tasks, it is assumed, will perform with similar effectiveness if they are given the opportunity to do so in the real world. This assumption has much common sense appeal which partially accounts for this method's growth in industry. The attempt to predict complex leadership and management behavior through procedures that are directly performance-related – the essence of the assessment center approach – is, of course, the major strength of this technique. However, while such direct assessment procedures appear to be highly performance-related, they lack validity because of subjective rater reliabilities. Furthermore, direct performance observation and assessment techniques are time-consuming, labor intensive, costly and less amenable than other techniques to quantification and statistical analysis. In general, however, behavior sampling techniques can be of great value if care is taken to assure their objectivity and reliability. (For proper critiques of assessment centers, see Dodd, 1970; Huck, 1973; MacKinnon, 1975).

Another type of simulation, which has drawn considerable attention in the medical profession, is the programmed examination (Hubbard, Levit, Schumacher & Schnabel, 1965; McGuire, 1963, 1969; McGuire & Babbot, 1967). This technique is used to measure patient management abilities of physicians. A simulated encounter between physician and patient requires the examinee to respond to patient complaint in a form and language that might occur in a real life situation. Alternative actions are printed in an answer book where the examinee indicates his/her response by erasing a special ink coating which conceals further information about the patient related to the response made. The examinee processes and analyzes new information after each response, making management decisions and receiving feedback. Sequential decisions and analyses continue through a set of problems related to the course of the illness simulated.

This technique has several advantages. It is less labor intensive and time-consuming than an assessment center. It is more objective and amenable to computer administration and scoring which permits more rigorous statistical and empirical analyses. It is particularly appealing in situations where assessment under actual conditions might prove hazardous to life. However, while this technique allows for some processing and integration of information, its format is

quite different from real practice. Furthermore, Hecht (1976) found that this technique had not been validated in relation to predicting job performance. Rather, what validation that has been done relates these test scores to medical school course work. This is disappointing in light of evidence that course work in medical school shows little relationship to competence in the profession (Price, et al., 1971).

Work Correlated Measures

The instruments in this category include paper and pencil tests which measure psychological constructs. From the test scores of those being assessed, assumptions or predictions are made about how one might perform in a variety of situations. These tests range from those that try to predict specific behaviors in limited situations to broad trait measures which purportedly tap some enduring attributes of personality or character that prevail in at least all normal situations. A proliferation of examples could be used here since assessment technology from earliest years has focused most heavily on correlational techniques. Intelligence tests, personality measures, and standard attitude, aptitude and achievement tests are the most common examples of this type of measure. They are developed by having a panel of experts (1) define the domain of content related to the area to be assessed, (2) choose the instances that it wants to measure, (3) sample appropriately from these instances, and (4) actually construct items or elements which represent them (Hoffman, 1976). A subset of this final group of items is empirically retained such that the items that remain differentiate between groups. Often the groups are made up of those who have high over-all test scores prior to performing an item analysis. This procedure is entirely tautological because it does not assess the validity (i.e., the job-relatedness) of the items designated by experts in the first place. Thus, it is not surprising that in general the correlation between tests of this type and performance criteria, though statistically significant, account for very little predictive power (e.g., a typical correlation coefficient of .30 translates to only nine percent of real predictability).

Indirect measures, including those just mentioned, often have high performer and rater reliabilities. They also tend to be efficient, objective, inexpensive and highly amenable to statistical analysis and treatment. However, they often lack validity because they are vague or unrelated to (unpredictive of) actual performance. Ghiselli (1965) conducted an exhaustive review of predictive studies for an impressively wide variety of jobs as a whole; the average of the maximal

performance predictive validity coefficients was a meager .33. Conversely, taking all tests' categories as a whole, the highest grand average performance predictive validity coefficient was .30. Matching the right test battery with the right job enhances these averages, but not impressively.

Another variety of work-correlated measures that is worth mentioning is the test of a construct that may mediate performance. These constructs, such as cognitive, interpersonal, motivational and other characteristics discussed earlier, are intuitively obvious as critical to success in professional roles, but they are rarely measured if at all. The hallmark of this category of measures is that they tap the types of covert antecedent causally-related variables reflected in the latent content of job and person analyses.

Clearly, the use of multiple-choice tests dealing with the subject matter of an occupation is untenable as a mechanism to protect the public interest. As we have seen, these measures cannot reflect even minimal standards of practice because they embrace only very limited aspects of competence. Nonetheless, licensing boards, certification agencies and professional test developers continue to rely on these measures and in many cases without the slightest evidence of their validity or meaning. Social pressures for accountability, control of manpower supplies or the desire for professional recognition have led to the proliferation of licensing without proper research from which to make reasoned decisions about entrance requirements.

Conclusion

The nature of competence is a complex phenomenon that has only begun to be properly researched and understood. We know that human performance is characterized by knowledge, skills, abilities and other characteristics which are only partially known and poorly measured. Efforts to simplify the richness and complexity of competent professional practices has only led us to presumptuously and self-righteously offer simplistic mechanisms for determining who should be allowed to practice within a profession or occupation.

Those who view behavior in its rightful context and who have developed better methods and techniques for identifying, defining and measuring aspects of competence have had little influence on those

who seek easy answers to complex problems. And some professionals within the testing industry are too ready to offer quick but inadequate solutions even when they know better and are capable of providing higher quality products to licensing agents.

More research is needed and should be supported by government agencies, foundations, the professions, the public and the testing industry itself. Guidelines for systematic planning and formulation of strategies to cope adequately with the enormous problems inherent in proper measurement should be enthusiastically and responsibly developed. Guidelines for utilizing better techniques already known as well as those that will become known should be made available to licensing boards and agencies. The high standards reflected in these guidelines should be rigorously enforced. Until such procedures for improving the state of the science and practice of assessment are implemented, a moratorium on new licenses should be legislated and enforced and interim procedures utilized which are fair to the professions, new candidates and consumers.

Chapter IV

Standards

The methods by which individuals and programs set standards for competent performance has been discussed at each of the meetings sponsored by the Task Force. The three papers which address that issue in this section of the monograph represent a wide range of decision rules. The range of rules represent, not only what was presented at the annual meetings, but also the current state of the art in the setting of standards. The Jaeger paper presents a methodological view of standards setting and warns the reader about the possible problems of validity. The Shepard paper presents standards in the context of human judgment and contrasts normative cut-off scores with the concept of competence. The Conaway paper reviews several procedures which are current practice in the arena of standard setting and discusses the consequences of each of those procedures.

The reader will not find a little jiffy formula in this section which will allow for the setting of standards. The choice of a standard is a complex decision. Each of the papers addresses the several factors which should be considered before a standard is set. Additionally, the ramifications of several current procedures for setting cut-off scores are discussed. While these papers were initially written in 1976 or 1977, we believe that they still represent current thinking on the issue of standard setting. Other material can be found in the Winter, 1978 issue of the *Journal of Educational Measurement* which has been devoted to the topic of setting standards.

Measurement Consequences of Selected Standard-Setting Models

Richard M. Jaeger
Univeristy of North Carolina at Greensboro

The early discussions underlying this paper led to a suggestion that I consider four models for standard-setting: a "regression model," a "professional judgment model," an "externally imposed standards model" and an "externally validated model." By title, all of these sound plausible. However, the literature provided no clues to definition or application. It seemed to me that there really were not four distinct classes of standard-setting models, only two: "judgmental models" and "empirical models."

Considerably more reflection and a more strenuous attack on the literature of criterion-referenced testing and competency-based applications of measurement convinced me that my initial categorization of models is without merit. All standard-setting is judgmental. No amount of data collection, data analysis, and model building can replace the ultimate judgmental act of deciding which levels of performance are meritorious or acceptable and which are unacceptable or inadequate. All that varies is the proximity of the judgment-determining data to the original performance. What I had earlier labeled a "judgmental model" I would now call a "proximal model" or "direct model," and what I earlier labeled an "empirical model" I would now call a "distal model" or a "derived model." In either case, subjective judgment of merit is inescapable.

One can hardly consider the act of forming judgments apart from the use of those judgments. Millman (1973), in a discussion of alternative procedures for setting performance standards, assumes that the ultimate purpose of judgment is to determine whether an individual could, if given the opportunity, successfully complete some prescribed percentage of the tasks in a well-described domain. Assessment of performance on the tasks in a domain is assumed to be an end in itself. Millman also restricts consideration of standard-setting to situations in which attribute performance, rather than variable performance, is of interest. The datum of interest is whether an individual could succeed in completing a given percentage of tasks; there is seemingly no interest in the actual difference between an individual's performance and the standard that defines minimally acceptable performance.

In this paper I will, likewise, restrict my remarks to attribute performance models. However, I will venture beyond the point where assessment of domain performance is an end in itself. The restriction is not dictated by my judgment of the ultimate utility or potential interest of variables models, only by space limitations. Consideration of purposes that go beyond assessment of domain performance appears to be an important and practical step. Suppose that you could administer all of the tasks in some well-defined domain and observe the result. For example, suppose that you knew with certainty that a given pupil could solve 72.9 percent of all two-digit addition problems. Or, to borrow an example from Ebel (1962), suppose that you knew with certainty that a given pupil could correctly recognize the definitions of 45.2 percent of the words in Webster's *New Collegiate Dictionary*. Would certain knowledge of these facts be useful? I strongly doubt it. Neither statistic would provide a prescriptive basis for remedial education. After all, if it's worth learning to solve any two-digit addition problems, it's worth learning to solve them all. More work is needed to raise the 72.9 percent to 100 percent, true. But which problems can the student solve now? Why were 27.1 percent missed? The statistic provides no answers or clues. Is recognition vocabulary of 45.2 percent of the words in Webster's *New Collegiate Dictionary* adequate? Is this a valuable finding? I suspect that a judgment on adequacy depends on tacit inferences to other domains: If recognition vocabulary is good, isn't generative vocabulary more likely to be good? If recognition vocabulary is as high as 40 percent of Webster's words, isn't sight reading likely to be satisfactory too? And won't this enhance school performance in the upper-level grades? And isn't reading skill necessary for success in professional occupations? Clearly, there are hundreds of potentially valuable (and possibly necessary) inferences of ultimate interest. Each inference depends on critical assumptions about the relationships between domain performance and behaviors external to the domain. Psychometrically, these assumptions are just as vulnerable as any other component of performance assessment.

Considerations and Purposes

Cronbach (1971) indicates that validity is not an attribute of a measure, but is associated with various inferences that are based on the results of measurement. A given measurement result, obtained under certain conditions and for certain populations, validly supports some

inferences and does not validly support others. Certain threats to validity arise because of the measure itself, others because of the way the measure is used, and still others because of the inference to be made. So it is with standard-setting models.

In this paper, I will briefly consider the use of a number of standard-setting procedures in support of two kinds of inferences: inferences to the performance of individuals on a well-prescribed domain of tasks, and inferences to the performance of individuals on some "ultimate" criteria that lie outside a sampled domain. For each procedure in each application, I will attempt to identify the principal threats to the validity of the desired inference. I will describe the nature of the threats and offer some conjectures on their seriousness. Discussion of the seriousness of threats to validity must be speculative, because the needed research is yet to be done. Perhaps this paper provides its best service by identifying a body of needed research.

Models and Problems

Proximal Standard-Setting Models

Inferences to Domain Performance: Where proximal data are used to set a performance standard and the inference of interest is to performance on a domain from which assessment tasks have been sampled, several procedures can be conceived. For each it is assumed that a judgment has been made on the percentage of tasks in the domain that must be completed successfully in order to label an individual "successful." This judgment could be based on an explicit definition of the kinds of tasks that compose the domain, such as an item form (Hively, Patterson & Page, 1968), or on examination of all of the tasks that compose the domain, as would be possible in Ebel's (1962) definition of domain-sampled tests.

Once adequate domain performance has been specified, there remains the practical problem of setting a standard for successful performance on a sample of tasks from that domain. Thus two judgmental acts are required to set standards in this situation – both domain behavior and sampled behavior must be considered.

Lord and Novick (1968), Hambleton and Novick (1973), Nedelsky (1954), Ebel (1974) and others have discussed ways of determining "mastery" scores once adequate domain performance has been defined. Some methods depend on the judged applicability of statis-

tical models and procedures; others depend on direct assessments of the value of individual tasks. Lord and Novick (1968) suggest that the maximum likelihood estimate of a domain score (percent successful) be used as a standard for assessing success on domain of tasks. The binomial distribution is used as a statistical model in this procedure. It is assumed that the set of tasks used to collect behavioral data is randomly sampled from the domain, and that the responses to these tasks are experimentally independent.

Let us now examine threats to the validity of the inference from the results of sampled behavior to the judgment of success on the domain of tasks. First, the standard of domain behavior that defines success was determined judgmentally. Given the same body of information on the nature of the domain of tasks, it is likely that different samples of judges — whether subject-matter experts, taxpayers, parents, or politicians — would set somewhat different standards. Thus the domain standard, given consistent definitional material, will vary randomly, and the judgment of an individual as successful or unsuccessful will likewise vary. The variance of judges' standards of success on various domains of tasks is largely unexplored, but would likely depend on the type of domain and the nature of the domain definition. Thus, there is a second source of judgment variance. How do judgments on standards for domain performance vary as a function of the definition of the domain? Suppose an item form is provided, or sample tasks are shown, or the entire domain of tasks is available for perusal. How do these factors influence standard-setting? Again, research is needed.

Once the domain performance standard is set, the validity of a maximum likelihood inference from a sample standard to the domain standard is threatened in at least two ways. First, the sample of performance might not be representative of domain performance; bias error may be present in the sampling of tasks. Second, the sample of tasks might not be of adequate size; too high a proportion of those judged successful on the sample of tasks might be below the passing standard on the domain, or too high a proportion judged unsuccessful might be above the passing standard. If the sampling of items conforms to the rules of simple random sampling (or those of a variety of other probability sampling procedures), the maximum likelihood estimator of percent successful on the domain of tasks will be unbiased. Millman (1972, 1973, 1974) has provided tables that aid in the determination of the number of tasks that must be sampled in

order to control the magnitude of random errors of estimation of a domain score.

Four sources of error then – random error among judges who set standards for domain performance, error due to the description of tasks in a domain, bias error due to inappropriate sampling of tasks, and random error due to an inadequate sample of tasks – threaten the validity of inferences from sampled behavior to domain performance in this case.

Novick, Lewis and Jackson (1973) propose the use of a Bayesian estimator of domain performance, based on the sampled performance of all individuals tested on a given occasion. All the threats to the validity of a standard based on maximum likelihood estimation would be present with this procedure as well. There are at least three additional error sources. First, a Bayesian performance of that individual's domain performance depends not only on the sampled performance of that individual, but on the performance of all of those assessed in the same group. Thus, using a Bayesian estimator, an individual could be judged successful when tested with one group, and be judged unsuccessful when tested with another group. Both bias errors and random errors in the selection of examinees are possible. The first may be due to a faulty sampling procedure or an inadequately defined examinee population, the last due to an examinee sample of insufficient size. Known theory (Novick & Jackson, 1974) can be used to determine the variability of Bayesian estimates of domain performance, as a function of the composition of examinee samples, examinee population definition, or examinee sampling procedures, but has not been explored in the context of competency-based assessment, and is another subject in need of investigation. It is known that Bayesian procedures tend to regress an individual's sampled score toward the mean of the distribution of performances of those in the sample group. The magnitude of the regression effect increases as the difference between the mean and the individual's performance score increases. Thus, an individual might well be judged unsuccessful when examined with a group of poor performers, but judged successful when examined with a group of high performers. Empirical research on the biasing effect of examinee selection on Bayesian success estimates is ripe for investigation.

Hambleton and Novick (1963) suggest that a decision on an examinee's performance, relative to a domain standard, should consider

the consequences of the decision, as well as a sample-based estimate of performance. Some cost or loss is incurred if persons whose true performance is above the domain standard are judged to be performing below that standard. Likewise, there is a loss (possibly different in amount) if persons whose true performance is below the domain standard are judged to be successful. The ratio of these losses as well as the probabilities of incurring each type of error, can be used to make decisions corresponding to each possible result of sampled performance so as to minimize expected loss. Either the binomial model based on individual performance, or the beta-binomial Bayesian model that uses individual and group performance, domain standard, given a sample of performance. If the assumptions of the beta-binomial model are appropriate, errors of misclassification will be smaller than those resulting from the binomial model. If the Bayesian prior distribution assumptions don't hold, even larger errors can occur. It would be interesting to conduct an experiment using actual performance data, together with performance histories, to check the correspondence between *a priori* distribution choices and actual distributions. Although the Bayesian approach has been shown to be theoretically robust by Novick and Jackson (1974) and others, the magnitudes of errors likely to be made in specifying *a priori* distributions are unknown. Threats to the validity of inferences from sampled performancé to domain performance identified for the binomial and Bayesian models above, apply with equal force to the decision-theoretic procedure when binomial or Bayesian estimation prodedures are used. In addition, judgments on the magnitudes of losses incurred as a result of errors are subject to variability across judges as well as bias errors that might be associated with different types of judges; e.g., teachers might have different perspectives on losses than would guidance counselors.

Several researchers have suggested that sampled tasks be examined individually and that judgments be made on such factors as (1) the probability that an examinee whose performance barely exceeds the domain standard could perform the task (Angoff, 1971), and (2) the value or worth of the task in estimating domain performance and the probability that a barely successful examinee could eliminate each incorrect option in a sample of multiple-choice items (Nedelsky, 1954). In each case, it is assumed that separate judgments of adequate domain performance have been made. For each of these methods of standard-setting, decisions on individual task difficulty are combined through an appropriate formula to establish a standard of performance on sampled tasks.

Inferences from sample performance to domain performance, based on methods that require examination of individual tasks, have the same validity threats as do those based on maximum likelihood estimation. In addition, the sample of judges used to evaluate individual tasks may be biased in their judgments. It is also likely that judgments of task difficulty will vary randomly across samples of judges. Thus, an examinee judged successful on the basis of one set of sampled judgments might be judged unsuccessful on the basis of another. The contribution of interjudge variability or systematic bias to the invalidity of such inferences is not known; again, research is needed.

Another class of standard-setting procedures required the collection and evaluation of performance data before the standard is established. In these procedures the need for judgment is not eliminated; the object of judgment is merely shifted.

In one such procedure, the purpose of evaluation is selection, and the object is to select those whose domain performance is better than that of some percent of a population of examinees. Thus, if the data were available, the domain performance of a population of examinees would be used to establish a domain standard. In practice, both examinees and tasks are sampled, and the distribution of performances of sampled examinees on sampled tasks is used to set a performance standard. Threats to the validity of inferences based on this procedure are numerous. The sampling of tasks and the sampling of examinees are subject to bias and random errors. Further, the judges used to set standards based on sample results may be systematically biased, and the variability of standards set by different samples of judges may be unacceptably high. Here too, empirical research is needed to establish the absolute and relative magnitudes of errors from these sources.

Inferences to Ultimate Criteria: Throughout the preceding discussion I have assumed that the inference of interest was to the proportion of tasks in a specified domain that an examinee could complete successfully, if given the opportunity. I have already argued that such inferences are often uninteresting or insufficient.

To make inferences that go beyond the domain sampled clearly involves additional threats to validity. First, one must question the appropriateness of the sampled domain: if one could observe perfor-

mance on the entire domain of two-digit addition items, would an assessment of this performance support an inference to performance on all addition problems? Would it support an inference to performance on problems involving all four basic arithmetic operations, provided only two-digit numbers were used? The questions are endless and the answers are few. Often our interest in current performance only substitutes for our true interest in later performance, perhaps years hence. To support inferences beyond the sampled domain of tasks, we must determine stable relationships between proximal and ultimate performance criteria. And these relationships must be realistic for a wide variety of examinee types assessed under many different conditions. Given our lack of success with seemingly elementary relationships in learning hierarchies, these threats to the validity of inferences based on standards of sampled performance appear serious indeed.

Distal Standard-Setting Models

Inferences to Domain Performance: A third set of standard-setting procedures involves the use of behavior in a related domain of tasks or on some external measure of performance to set the standard of performance for the domain of central interest. As an example of these procedures, suppose the domain of interest was composed of those tasks judged to be essential to successful classroom teaching. To observe performance on a sample of tasks for a large group of examinees would be costly and difficult. The use of actual schools would be required (a procedure that could be questioned on ethical bases) or complex simulations would have to be designed and conducted. As a less expensive surrogate for the domain of interest, one might be tempted to define a domain composed of pencil-and-paper tasks like those on the National Teacher Examination (NTE).

If a procedure analogous to the one described were used, a two-stage chain would be required to support inferences from observed performance to performance on the domain of interest. Three kinds of judgmental standards would be required. First, a standard of performance on the domain of ultimate interest would be needed. Second, a standard of performance on the sampled domain would be required. Finally, it would be necessary to set a standard of performance for observed behavior. The inferential chain would extend from the observed performance to the sampled domain, and then from the sampled domain to the domain of ultimate interest.

Threats to the validity of inferences based on the three types of standards abound. Inadequate descriptions of either or both domains of tasks could bias standard-setting. The samples of judges used to establish standards could be unrepresentative of an appropriate population of judges, resulting in bias error. The samples of judges used to establish standards could be inadequate in size, resulting in unacceptably large random errors. If relationships between performance on tasks from the sampled domain and those from the ultimate domain of interest were determined empirically (using samples of tasks and examinees sufficiently restricted in size to make the assessment feasible and affordable), bias and random sampling errors from three sources would threaten the validity of the findings. We have little information on the probable magnitude or seriousness of these threats.

Inferences to Ultimate Criteria: There is really nothing new to be said of these cases; the threats to validity of inferences already suggested for the simpler situations apply here as well. As might be expected, the validity of inferences from performances on a sample of tasks from one domain, with a standard inferred from a second domain, and desired inferences to yet a third domain (the essence of these cases), is subject to a great many risks. Enumeration of these risks would be repetitive. Just consider the threats to validity of inferences based on the maximum likelihood model, add those associated with inferences to ultimate criteria, and include those associated with the distal standard-setting models already described. The list is discouraging, if not mind-boggling, and our knowledge of the magnitude of errors and the severity of such validity threats is extremely limited.

An Attempt to Summarize and Some Final Judgments

Threats to the validity of inferences, based on various standard-setting models, are summarized in Table 1. The table is in matrix form with standard-setting models forming one dimension, and threats to the validity of inferences forming the second. Two classes of models are indicated – proximal and distal – although the utility of the distinction is open to question. Likewise, two classes of inferences are used to sort threats to validity. Threats to validity listed under "Inferences to ultimate criterion performance" should be considered supplements to, rather than substitutes for, those listed under "Inferences to domain performance." An "x" in the body of the table indicates that a threat in a given column applies to a model in a given row.

Table 1

Threats to the Validity of Inferences Based on Selected Standard-Setting Models

Standard-Setting Model and Procedure	Threats to Validity of Inference to Domain Performance										Additional Threats to Validity of Ultimate Criteria		
	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Judges set domain passing score directly a. Maximum likelihood estimate of domain passing score used as sample passing score	X	X	X	X							X	X	X
b. Bayesian estimate of domain passing score used as sample passing score	X	X	X	X				X	X		X	X	X
c. Decision-theoretic model (with losses) used to set sample passing score	X	X	X	X						X	X	X	X
d. Judges determine criterion worth of each task directly	X	X	X	X	X						X	X	X
2. Estimate of a percentile used to set passing score Tasks are administered to a criterion group; percentile is is used to set passing score	X	X	X	X		X	X				X	X	X
3. Performance in a related domain or on an external measure used to set passing score a. Contingency table analysis based on two domain samples of tasks (e.g., one costly, one inexpensive)	X	X	X	X				X	X		X	X	X
b. Mathematical model relating two domains or performance on domain tasks and an external measure	X	X	X	X				X	X		X	X	X

Key

Threats to Validity of Inferences to Domain Performance:

1 = Bias in setting domain standard due to inadequate domain definition
2 = Random error among judges who set domain standard
3 = Inappropriateness of item sampling procedure, bias error in sample standard
4 = Inadequate item sample size, random error in sample standard
5 = Judgement bias in the consideration of individual tasks
6 = Lack of representativeness in the criterion group used to determine percentiles
7 = Inadequately large criterion group, resulting in random error
8 = Lack of representativeness in examinee group, leading to bias error
9 = Inadequately large examinee group, leading to random error
10 = Model bias due to inaccuracy in loss judgments, false positives or false negatives overvalued

Additional Threats to Validity of Inferences to Ultimate Criteria:

11 = Bias error due to invalidity of domain identification
12 = Error due to inconsistency of domain criterion relationship
13 = Bias error due to invalidity of model for domain-criterion relationship

The research that exists on standard-setting procedures in competency-based education appears to be largely theoretical. Millman's work on the relationsnip between random errors and the size of task samples is based on a well-known binomial probability model and several plausible, but untested, procedural assumptions. Hambleton and Novick's results on the application of decision theory and beta-binomial analysis to performance estimation likewise makes use of elegant statistical theory.

Throughout this paper, I have identified questions for which research-based answers are apparently unavailable. Application of statistical models and theoretical formulation are unlikely to provide answers to these questions. There is need instead for empirical investigation involving human standard-setters in real or simulated judgmental situations, using real performance data and real descriptions of task domains. Perhaps a theory of validity will emerge from this research eventually, but tentative guides to practice, sorely needed, will be developed in the interim.

Setting Standards

Lorrie A. Shepard
University of Colorado

The avowed virtue of competence-based education is that instructional efforts are accompanied by unambiguous criteria for learning proficiency. Competence-based assessment requires standards to distinguish acceptable and unacceptable performances. The purpose of this paper is to consider practical constraints on standard-setting and to make recommendations. This paper is written to complement Dick Jaeger's paper in this monograph which treats the psychometric issues involved. It begins where Jaeger's leaves off with interim suggestions in the absence of empirically validated procedures.

Definitions and Distinctions

Before proceeding with practical suggestions for setting standards, some definition difficulties that cloud the issue of setting performance criteria ought to be cleared up. There are some terms that are appropriately synonyms for "standards" and others that are not. There are some concepts that seem perpetually linked with competence criteria that are really unrelated notions.

For example, performance standards are sometimes thought to be the same as normative cut-off scores. That is, criterion scores are set so that a prespecified number or percent of candidates pass. Millman (1973) pointed out that whether an individual succeeds by this type of criterion depends in part on the competence of others taking the test. While this procedure is appropriate when "the number of people to be certified is fixed" (Millman, 1973, p. 206), such a criterion lacks any intrinsic standard. It ensures that the most able candidates will be selected but does not guarantee the skill level of any of them. Correct use of the term "standard" in the context of this paper requires an absolute rather than a relative judgment of performance. With this stipulation understood, "criterion score" may be used interchangeably with "performance standard."

The purpose of standards is to identify mastery of whatever content is being assessed. Performance criteria or standards are a necessary feature of mastery learning of the types advocated by Carroll (1971)

and Bloom (1971a). The standard is the level of proficiency that each student is expected to attain. One of the difficulties with mastery learning is its basic tenet that given enough time nearly any student can learn what we wish to teach him. Depending on the difficulty of the material, this fundamental assumption may or may not be met. This issue should not, however, be confused with the equally perplexing problem which we are facing of how to determine the level of performance that constitutes mastery. This distinction is very important for subsequent discussion, since when these issues are mixed, standards must be set as minimums so that all students can attain them. In a recent article, Spady (1977) acknowledged the similarities between mastery learning and competency-based education. These two concepts may not be distinguishable, particularly when the objectives being assessed are basic skills. The term mastery implies a high percentage criterion score. Advocates of mastery learning (Block, 1971; Bloom, 1971a and 1971b) acknowledged that the concept is more appropriate for circumscribed subjects which do not require creative or divergent thinking. It is conceivable that for some complex learning tasks a low criterion score could be the standard for competency, i.e., less than complete mastery is acceptable. However, such distinctions between testing for mastery and competency-based measurement are likely to occur infrequently.

Another question which is often confused with the problem of setting standards is the method of test construction. The terms "objective-referenced testing," "criterion-referenced testing," and "domain-referenced testing" are often used as if they meant the same thing. The confusion of these terms in common parlance will probably never be eliminated, but the distinctions that have been obscured ought to be redrawn. Criterion-referenced testing (CRT) was meant to describe what students know or have accomplished rather than their relative standing in a group (Glaser, 1963). The name is now taken to imply that a passing score or criterion is established to determine which students have succeeded. Criterion-referenced testing was also characterized by a more careful specification of the content universe to be assessed. This systematic development of test content is a separate feature and is probably now better characterized by the term objective-referenced testing. Many so-called CRTs are objective-referenced or content-referenced but do not have standards of performance. The name implies that standards or criteria for judging acceptable performance are a part of the testing process. Instead, it

process. Instead, it may be that the test items have been developed to correspond to a detailed schema of the content, but that no standard has been agreed upon. Certainly setting criteria is partially dependent on properly referencing items to objectives. How meaningful would criteria be if the content universe were not carefully identified and appropriately represented by the items? However, it is important to realize that the methodological insights about how to develop objectives and subobjectives do not contain the solution to the problem of how to set criteria.

Domain-referenced testing, described by Millman (1972) and Hively (in Maxwell, Hively, Lundin, Rabehl, & Sension, 1971), is an even more precise method of test content specification. The item-forms which characterize domain-referenced testing are intended to be explicit enough to determine the difficulty level of all items generated from a particular form; in this sense, criterion levels are a part of each item form. Nevertheless, when performance is aggregated across items, there is still the problem of what combination of success and failure denotes acceptable performance.

Finally, the concept of competence-based assessment seems to be tangled with notions of learning hierarchies, prerequisite skills, minimum competence, basic skills and the more popular survival skills. Gagné (1965) is the principal exponent of a seductively simple theory of education. Every learning task is comprised of elements. Learning is facilitated when each task is analyzed into its requisite subtasks so that the learner may proceed sequentially through the implicit hierarchy. The theory is intriguing and has some empirical examples (Gagné and Paradise, 1961). It seems to work best in math where, just as in the case of domain-referenced testing, we are better able to analyze goals into specific subskills. Nevertheless, there are counter-examples where the theory does not work so well. Learning hierarchies are only absolute when the order of skill acquisition is immutable for all individuals. For most learning tasks, the hierarchies are not fixed or absolute. There are always anomalous individuals who have mastered an apparently complex skill without mastering supposed prerequisites: like playing the piano without learning to read music. In math, it may be that knowing how to add two-digit numbers is an absolute prerequisite for adding three-digit numbers; but it may also be that addition facts need not be mastered before learning to multiply. Observations like these pose a conundrum for standard-setters in competence-based assessment. If the purpose of

setting criterion levels on the earlier tasks is to ensure success on later endeavors, the counter-examples are very troubling indeed. What if a child has not mastered a prerequisite skill and is assigned additional practice? If the prerequisite skill has no merit in itself but is only being studied as the means to an end, then the extra practice may be a waste of time. If some children can master the final objective without the prerequisites, then we are less certain of our standards for mastery.

The learning hierarchies model is not very practical for setting standards. It is useful in only very circumscribed curricular areas. When the ultimate tasks are broad such as "being a good citizen" or "being successful in life," it is very unlikely that valid hierarchies exist. Standard-setters would be better able to fulfill their role if they did not have the false scientific sense that such hierarchies exist in nature if they could only develop the methodologies to uncover them. "Standards are choices, not essences."

Setting Standards Requires Judgment

If in all the instances that we care about there is no external truth, no set of minimum competencies that are necessary and sufficient for life success, then all standard-setting is judgmental. Our empirical methods may facilitate judgment making, but they cannot be used to ferret out standards as if they existed independently of human opinions and values.

So-called empirical methods for standard-setting usually involve statistical means for maximizing the prediction of some ultimate task from the competency assessment. For example, a cut-off score may be set on a medical certification exam to maximize the agreement between success and failures on the exam and success and failures in the profession. Nevertheless, judgment is integrally involved in the definition of the criterion performance. Although the cut-off score is apparently determined statistically, judges actually decide it depending on the level of success chosen on the external criterion and whether professional success is defined as annual income or patient longevity.

The subjective nature of standards is awkward when their result is interpreted with black-and-white certainty. Glass (1977)* examined the various methods for setting standards. Each, he demonstrated, is arbitrary or relies on arbitrary premises. For example, Glass calls one method"counting backwards from 100 percent." The desired criterion is 100 percent, but to be realistic, standard-setters allow for ordinary mistakes that are not inconsistent with the concept of competence. However, "just how great a concession is to be made becomes distressingly arbitrary with some allowing a five percent shortfall and others allowing 20 percent or more" (p. 16). Sophisticated techniques for determining cut-off scores only mask the essential arbitrariness of the task. Glass elaborated on a study by Andrews and Hecht (1976). They conducted an empirical comparison of the Nedelsky (1954) and Ebel (1972) strategies for determining minimum competency levels. By the Ebel method, 69 percent of the items must have been answered correctly to certify competence; by the Nedelsky method, only 46 percent was required. The seriousness of this difference in criteria is illustrated by Glass's estimation that by Ebel's standards only seven percent of the examinees would pass and by Nedelsky's standard 63 percent would pass. Glass concluded that:

> The impression of scientific objectivity created by the rigmarole of grids and guessing corrections quickly evaporates when one sees the staggering discrepancy between the pass-rates of the two standard setting methods. (p. 26)

There may be some lingering hope that the relative error among standards set by the same method would not be as great as the discrepancy between these two standards set by different methods. But the seeds of doubt have been sown; the possibility for mistakes is apparently great; and we have no way of telling when we have established a valid standard and when we have not.

Errors in single standards are compounded when a profile of pass and fail scores is interpreted. Apparent strengths and weaknesses may be incorrect for one of two reasons. First, in setting standards for a dozen objectives, judges may have varied the stringency of their expectations. What may seem to be an area of greatest deficiency

*Editors note: This paper was reprinted as the lead article in a special issue of the *Journal of Educational Measurement* Volume 15, Number 4 Winter, 1978 devoted to setting standards.

may actually be an objective for which the standard was set unrealistically high. At the same time, shoddy performance in another area may go unnoticed because of an overly lenient standard. The second cause of misunderstanding is differences in the difficulty of items used to assess each objective. It is easy to imagine that for different sets of items (or ages of examinees) that a 90 percent criterion could be either trivially simple or hopelessly out of reach. If this is as it seems, why are there so many examples of whole sets of objectives with their criterion levels fixed at a constant percentage? It is absurd for the standard on every objective to be 85 percent unless objectives are equivalent in importance and item difficulty. Even if judges try to compensate for item difficulty by adjusting the standards, it is possible that unevenness in the harshness of standards will be mistaken for valid performance differences. Millman (1973) recommended that the problem of variability in item difficulty can be forestalled by the use of item-forms and more careful domain-referenced testing. Unfortunately, this kind of precision in item construction has only been demonstrated for mathematics.

Expert Judgment is Essentially Normative

Having said that all standard-setting requires subjective judgment, it is important to make the seemingly contradictory point that expert opinions are implicitly normative. When judges are asked to set writing standards for sixth-graders, their decisions will be based on experience. But what do we mean by "experience" except imperfect norms? Each expert is likely to base his opinion on his own sixth-grade experience, or that of children he has taught or that of his own sons and daughters. For standards to be realistic or attainable, experts are likely to reflect on what current sixth-graders are able to do; by relying on experience, their standards will be normative but will lack the representativeness of more systematic sampling. This apparent heresy, which is likely to be an anathema to devout criterion-referenced testers, ought to be acknowledged and is an important part of the recommendations for standard-setting. There are some advantages to judges being influenced by formal and explicit norms rather than unsystematic and internal ones.

Standards should not be the Lowest Common Denominator

Those who believe that standards exist external to the judgment of experts are likely to engage in searches for the essential skills. They

might, for example , interview plumbers and shop clerks to try and locate which competencies are minimal, and which are held by all employed adults. Such searches will be informative and may be helpful to judges in the same way that normative data are helpful, but the searchers will not turn up any universals. We might as well consider right now what we will do with some very successful plumbers who cannot read or high-paid street sweepers in San Francisco who cannot handle simple fractions. The search for absolutes leads to absurd reductionism. Perhaps we should study the mentally retarded. What skills are absolutely essential to be able to ride the bus to and from a sheltered workshop? Suppose we thus identify some basic skills that are completely rudimentary. How would these lowest-common-denominator standards serve as meaningful criteria for prospective carpenters and TV repairmen?

As I remarked earlier, Gagné's learning hierarchies have had a pervasive effect. Belief in his tenets has sent assessors and standard-setters in search of criterion scores as if they were external truths. In addition, an overly simplified view of learning has been adopted. Although any student may demonstrate mastery with lots of coaching and prompting, what disciples of mastery learning don't consider is how unequal students may look on retention and transfer. Instead of recognizing that students may reach different levels of mastery, adherents are likely to seek more and more fundamental tasks so that evidence of success and failure will be black and white instead of gray. Those who believe that standards ought to be the lowest common denominator have failed to recognize that the behaviors we wish to predict, and the life skills we wish to ensure are not singular but are arrayed along a continuum. The competencies that mentally retarded individuals ought to have to lead a full life are not the same as those necessary for other subgroups to lead rich and productive lives.

Judgmental Standard-Setting is Subjective but not Capricious

The theme of the preceding sections is that setting standard is judgmental. The validity of the standards will depend on the wisdom of the standard-setters. No standard inheres in nature for them to discover. If nothing else, standards will be established more sensibly and with less argument and grief if experts understand their limitations and their role. There are no magic ratios, 80 percent or 90 percent, that can be arbitrarily assigned regardless of skill level or content domain.

This recommendation seems so obvious, but it is the single most frequent error in criterion-referenced assessment. In the evaluation of the Michigan assessment results, House, Rivers, and Stufflebeam (1975) criticized the use of the term "minimal" for objectives in math and reading that were not attained by any of the districts in the state. Since some of these districts routinely scored very high on traditional achievement tests, the fault seemed to be with the objectives rather than with the educational system. In retrospect, it appears that not much thought was given to the use of the word "minimal"; either the assessment should have been conducted with the existing objectives but without the minimal designation or at least better judgment should have been made about which objectives were truly minimal. Obviously, identifying minimums is not easy but experts should ask themselves which objectives are so essential that districts can be considered negligent when the necessary percentage of students does not attain the objective. If experts had asked themselves this question in the Michigan case, they might have forestalled the embarrassing results.

The Harshness of Standards Ought to be Modified Depending Upon the Seriousness of Two Types of Errors

If standards were set wisely, it is assumed that most decisions that result from the assessment process would be correct. Individuals who were incompetent would fail the test and those who were competent would pass. Mistakes will occur, however, because of measurement error or improper inferences to the real-life skills for which the test is a proxy. These errors are of two types, viz., false positives and false negatives. False positives are those who pass but do not have the necessary mastery; false negatives are those who fail but actually possess the requisite skills.

The seriousness of these two kinds of mistakes will vary with the situation. When individuals are certified to practice in various professions as doctors, lawyers, teachers, the cost to society is much greater for the false positives than for the false negatives. In these cases, relatively stringent standards ought to be set to protect the public against unqualified practitioners. The cost to individuals who are thereby unfairly failed is outweighed by the public good. In many instructional settings, however, the reverse is true. In instances when learning hierarchies are valid, strict criteria prevent an individual from encountering material that is too difficult until the prerequisite

skills ahve been mastered. But when learning hierarchies are not accurate, it may be that the most serious costs occur with the false negatives. Individuals who are forced to drill on material they have already mastered by become frustrated or bored. Such a circumstance is particularly untenable when incomplete mastery of early skills does not actually prevent mastery of more complex skills.

Expert judges will have to consider the relative costs of the two types of error and adjust the standards to protect against the most serious mistakes. The current cry for competence-based high school diplomas is motivated entirely by one type of error – those who graduate without seemingly essential competencies. Before these various plans and laws are enacted, someone ought to raise the issue of the other type of error. Are there societal or personal benefits that would warrant keeping some nineteen-year-olds in school? Will we be better off if diplomas are withheld from those who do not meet certain standards? This question has not been raised because in the current rush for minimum standards, most believe that the existence of standards will increase the skills of high school graduates. Maybe they will. But if standards are to be meaningful, there are bound to be individuals who cannot meet them. Some will take a longer time, others will never make it. Only trivially low standards will be passed by everyone.

Recommendations for Setting Standards

1. Avoid setting standards whenever possible.
The remaining suggestions are intended to make standard-setting as thoughtful and well informed as possible. But these precautions cannot compensate for the errors implicit anytime a continuum of performance or skill development is represented as a black and white dichotomy. Therefore, pupil progress should be monitored along a performance continuum which reflects degree of accomplishment. Mastery classifications should be made only when discrete categorizations are forced by instructional constraints such as the need to assign children to remedial groups. Even then cut-off scores should be set with an eye to the number of children who can be accommodated; and the fallibility of the placement criterion should be acknowledged by providing easy access to retesting.

As has already been discussed, errors in single standards become more serious when a profile is interpreted for several objectives or several subject areas. Apparent strengths and weaknesses may only be differences in the stringency of the standards. This problem exists whether one is evaluating the attainments of individuals or school programs. In both instances, one can have greater assurance that strengths and weaknesses are real if relative deviations from norms (or norms adjusted for background characteristics) are considered rather than pass-fail results on judgmentally set standards.

2. Make setting standards an iterative process.
Subsequent recommendations suggest ways to enlighten the judgment-rendering process; but because it relies fundamentally on human wisdom, there are bound to be errors. Thus, criteria should not be fixed for all time. Panels of experts ought to be reconvened when testing results are in. They should examine the results, looking especially for examples of false positives and false negatives. The purpose of their review should not be to grant exceptional certifications of mastery, but to find systematic errors that suggest a change in criteria is needed. They will be conducting an informal validation, looking for agreement between the results of criterion assessment and additional data and judgments. In the Michigan example, experts might decide to adjust the standards if all school districts failed unless they believed that all districts were indeed failing.

3. Include the normative basis of judgments as a formal part of the standard-setting process.
Expert judges ought to be provided with normative data in their deliberations. Instead of relying on their experience, which may have been with unusual students or professionals, experts ought to have access to representative norms. If new measures are being developed along with the standards, then normative information could be obtained from similar existing measures or from field trials.

Of course, the norms are not automatically the standards. Experts will still want to decide what "ought" to be; but they can establish more reasonable expectations if they know what current performance is than if they deliberate in a vacuum.

4. Adopt "improvement" as the most reasonable standard.
In some areas, it may be possible to establish minimal competencies that are absolute and consensual, where everyone agrees that a skill

is essential and mastery is clearly distinguishable from non-mastery. For example, all physicians ought to be able to recognize and treat shock. However, in most areas of education and training, these absolutes do not exist. In these instances, "improvement" may be the only defensible standard. Having looked at data on current performance for a state or a district, experts ought to identify those objectives for which performance has been satisfactory and those which need improvement. There is still the difficulty of deciding how much improvement is needed, of actually setting the standard. But, designating areas where improvement is needed is an important first step. If the standard-setting process is really iterative, the standard will evolve over repeated assessments. It will be "set" when the experts decide for a particular objective that the performance level is acceptable and ought to be "maintained" and that attention for improvement be focused elsewhere.

National Assessment was intended to provide information about what students and adults in the United States know and can do. It was the creation of those who advocated criterion-referenced testing, but standards were not specified. There was very little excitement over the result of the first assessment rounds because most audiences were unable to judge for themselves whether the reported percentages were good or bad. The standards were missing. Now there is a great deal of excitement because the science and writing results have shown decline from the first to the second assessment. After verifying that the science decline could not be explained by a change in objectives or a curricular shift from the physical to the biological sciences, it was not difficult to assign meaning to the results. A decline is bad.

5. Allow for differences of opinion by involving various audiences in standard-setting.

Jaeger, in the previous paper, points out that the validity of standards depends on the sampling of judges. This is true whether the judges establish the standards directly by assigning criterion scores or indirectly by defining success in a criterion group.

It is a fairly obvious and hackneyed recommendation that all relevant audiences ought to be represented in the setting of standards. It is worth mentioning, however, since in some instances where legislators or militant parents have insisted on the setting of standards, they have left the actual judgments to professional educators. This is per-

plexing since there is no reason to believe that the educators will reflect the values held by parents and taxpayers. The impetus for competence-based assessment in the first place may have been differences in values and priorities. If an organization of employers has lobbied for competence-based diplomas because high school graduates cannot make change, their values ought to be represented in the setting of standards. It is only postponing the clash if educators substitute standards in the affective domain.

Representation from groups who disagree may be the most straightforward way of dealing with differences in values. Brickell (1974) wrote a very amusing paper about the external factors that influence evaluation. The evaluator was buffeted by enormous political pressures in all settings but one. The fairest and most politically neutral situation for the evaluator was that in which all sides were represented.

Once experts have been identified from relevant audiences, they should not be tossed together to reach consensus on standards. Obviously, consensus standards would be easier to implement. But, personality dynamics in a particular group of judges could create phoney consensus. The best protection against artificial consensus is to convene more than one group of expert judges and have them meet separately. If they arrive at the same standards or nearly the same standards independently, their agreement will be observable and will be much more dependable than if they had met jointly.

If groups of experts do not agree on what standards should be, alternatives still exist for arriving at standards. If standards are being insisted upon as a means to evaluate school systems, it is preferable to report assessment results in light of more than one criterion. This is similar to the apples and oranges problem encountered in comparative evaluation. One program may look better by one set of criteria; the other may shine when judged by alternative criteria. The only fair and comprehensive way to judge them is to apply multiple criteria to each.

If the standards are to be used to make decisions about individuals, then one set of criteria is needed. The criterion could either be the most stringent or the most lenient of those proposed depending on which type of error is more serious in that situation. If false negatives would be more costly to individuals and society, then the standards

should be lower than in the instances when false positives must be screened out.

6. Protect against focusing on minimums which may limit the height of educational attainment.

Competence-based education certainly does not imply that learning will stop after the basics have been mastered. It only implies that learning is sequential and that mastery of early tasks will be accomplished before passing on to the next. However, the current popularity of this concept with legislators and the public carries with it the term minimum. Certainly for individuals, there cannot be much quarrel with learning the basics before attempting more complex tasks. But for entire school systems, there may be some unforeseen consequences if the exclusive focus becomes the attainment of minimums. Ultimately, we will have to face the choice of whether to teach the last student in a school to work with decimals instead of teaching a classmate the beginnings of biochemistry. Once competence-based assessments are fully installed, parents and taxpayers ought to ask for assurances that minimums are not being attained at the expense of excellence.

Anti-testers are fond of the argument that testing minimums will limit educational growth?. Rather than being a plea for less testing, however, this caution warns of the need for collecting evidence at both ends of the performance continuum.

For individuals as well, there is the concern that required minimums will become the maximums, that students will stop trying when minimums are attained. Part of the solution to this problem is the observation provided earlier that the life skills that are ensured by competence-based assessment are arrayed along a continuum. A full and productive life for most individuals requires more than the minimums. Perhaps assessment and rewards for accomplishment beyond the minimums are means for increasing growth towards these goals.

Setting Standards in Competency-Based Education: Some Current Practices and Concerns

Larry E. Conaway
Research Triangle Institute

One important component of competency-based measurement is the setting of standards to be used in determining if students have attained adequate proficiency to advance to the next step in any given process. A review of the literature shows that a definitive set of practical procedures for setting standards in various types of competency-based programs does not exist. Airasian and Madaus (1972) stated that the area of setting standards was the area of criterion-referenced measurement most in need of research, and Quirk (1974) discussed a similar lack of research in the context of performance-based teacher education. Although increased attention has been given to the issue in the past few years, it is apparent that practitioners cannot obtain validated standard-setting procedures, either off-the-shelf or out of the literature. Instead, they must devise their own standard-setting procedures, relying upon their own experiences and philosophies in competency-based measurement and upon any guidance they can get from available literature and fellow practitioners.

Purpose and Scope

The purpose of this paper is to discuss standard-setting from the perspective of the practitioners who are presently faced with implementing competency-based programs. The literature is reviewed (1) to provide an overview of the status of standard-setting and (2) to determine, in conjunction with personal conversations with practitioners, what procedures are being used to set passing standards for some competency-based programs which are presently being implemented or designed. Based upon these foundations the state of the art is then discussed in terms of several factors that are considered to be of importance in setting standards of proficiency.

In the broad sense competency-based education covers many different types of programs. The following types of programs would certainly be included:

1. proficiency tests providing for early graduation from high school;

2. proficiency tests which must be passed in order to graduate from high school or be promoted to the next grade; and

3. various applications of criterion-referenced tests in curriculum and instruction.

Within each type of program there are a great number of variations. For example, proficiency tests for early graduation from high school might be mandated by either the state or local district, might focus upon either basic skills or life applications, and might be the sole criterion for graduation or one component in a set of criteria.

For the purpose of this paper competency-based education will not be narrowly defined. Models and examples have been drawn from each of the types of programs, and there are many more similarities than differences in the problems associated with setting standards within each type of program. While some practitioners may find that some information is irrelevant to their particular situation, a more common problem will be that no definitive information is available from any easily accessible source.

A Note of Caution

Those of us who have extensive backgrounds in educational assessment, which must be considered one of several precursors of competency-based educational programs, are generally surprised by the direct use of test results in making important decisions such as granting early graduation, withholding diplomas and establishing remedial education programs. Several assessment programs at the national, state and local levels have generated results of excellent quality and been unable to have any major effects on important decisions about programs and students. In many cases this has been true in spite of strong efforts to involve members of important groups, to generate and disseminate interpretations of results, and to establish standards of performance.

For most of those who have been involved in attempting to get assessment results utilized, this has been a truly frustrating experience; therefore, at first glance the direct use of test results in making important educational decisions is extremely encouraging and

refreshing. However, upon more careful reflection about the important and possible implications of some of the decisions being made and about the unresolved issues in test development and measurement, one begins to find the initial encouragement being balanced by feelings of apprehension.

The present movement into operational programs is probably inevitable and timely given the educational focus of the last decade. Clark and Thomson (1976) correctly stated that the use of competency tests is a logical step based upon two parallel concerns in recent years: (1) the educators' concern with behavioral objectives, student performance outcomes, and criterion-referenced tests; and (2) the public's critical review and analysis of student performance. However, even though educators have been involved with the measurement issues of criterion-referenced testing for over a decade, it is still true that unless we move with caution and discretion, we will quickly move ahead of the state of the art. In discussing competency-based education Spady (1977) emphasized that although the problems associated with curriculum and instruction are great, they are minor compared with those related to reliable and valid measurement, whose technology is in its infancy.

The Literature

Standard-Setting Models

In addition to the papers by Jaeger and Shepard included in this monograph, there have been some other comprehensive reviews of standard-setting models and their applications.

Millman (1973) discussed seven different procedures for setting a passing score organized around five factors: performance of others, item content, educational consequences, psychological and financial costs, and errors due to guessing. Millman discussed the assumptions associated with the various procedures; and although he did not recommend a particular procedure for setting passing standards, he did indicate testing conditions under which each procedure would be particularly appropriate or inappropriate. In this work Millman also discussed two issues that are closely related to the passing standard: estimating the domain score, and test length. In a later work Millman (1974) discussed these two related issues in greater detail.

Millman's review was set in the context of domain-referenced testing

while Meskauskas (1976) reviewed standard-setting procedures in the context of evaluation. Meskauskas began with the following statement:

> The adoption of the criterion-referenced approach to evaluation very quickly raises two measurement issues that have relatively less importance in norm-referenced testing. These can be broadly stated as the issue of the definition of mastery, and the issue of a priori standards – closely intertwined but different problems. Rigorous exploration of these to date has been quite minimal. Perhaps the development of content aspects has been of greater urgency. However, this area is receiving increasing attention. Measurement specialists have turned their attention to criterion-referenced measurement, introducing the use of decision theory and Bayesian statistics. (p. 133)

Meskauskas organized his discussion of various standard-setting models around two different views of mastery: continuum models in which mastery is viewed as a continuously-distributed set of abilities, and state models in which mastery is viewed as an all-or-none task. He also discussed some standard-setting models which do not fit either of the views of mastery; these models deal with the problem of test length, which is an issue closely related to standard-setting. Within this structure Meskauskas discussed various methods which have been developed for establishing minimal proficiency levels and determining if students have attained them. He discussed the underlying assumptions of the various models and some of the strengths and weaknesses associated with their design and application. Meskauskas concluded that no standard-setting models have been designed and validated for general use. He added that measurement specialists must validate their standard-setting methods and provide users with data on which to make choices, and he cautioned that decision procedures must take curriculum processes into account and be practical for use by instructional personnel.

Competency-Based Programs

While progress has been gradual in establishing validated standard-setting procedures and conducting research into the surrounding measurement issues, operational programs are being planned and implemented at an ever increasing rate. Spady (1977) dates the beginning of the movement as 1972 when the Oregon State Board of Education passed its first regulations regarding minimal competencies for graduation. As with most popular movements, or bandwagons, similar thinking had been taking place in other geographical areas; and

most states and many localities have taken steps to plan and implement competency-based programs since that time.

Another indication of the increasing popularity of competency-based programs is the development of commercially available criterion-referenced tests. Kosecoff and Fink (1976) investigated the feasibility of using criterion-referenced tests in large-scale effectiveness evaluation; and as a part of this investigation, they requested criterion-referenced reading or mathematics tests from test publishers. They received 28 tests that were potentially useful and had accompanying test manuals, and these were from 12 different publishers. If a long run general demand for such tests is perceived by test publishers, the investment will be made, just as it is presently being made in the area of applied performance learning by some test publishers.

The Status

Practitioners can certainly obtain some information relevant to their standard-setting problems from the literature; however, they should not expect to find ready-to-use procedures that have been validated or undergone extensive use. Millman and Meskauskas each acknowledged that most passing scores have been set arbitrarily in the past, based upon such things as tradition and the intuition that the chosen score represents mastery but does not demand perfection. Kosecoff and Fink also found that most of the commercial criterion-referenced tests they reviewed either left the passing score decision to the test user or reported "mastery" and "non-mastery" scores based upon correctly answering an arbitrarily selected number of items per objective. At the present time there is no indication that the literature has had a great impact on practitioners, and conversely there is no indication that the practitioners have made major contributions to the literature.

Standard-Setting Practices

In the remainder of this paper some standard-setting procedures that are currently being implemented will be described, and some of the problems and issues being faced by practitioners will be discussed. The discussion includes some explicit and implicit recommendations, or advocacy position. Although there is a desire that some of these will be helpful to specific practitioners, it is realized that the many competency-based programs with their diverse purposes and designs will create an array of standard-setting situations, which will be

approached in a variety of ways by practitioners with different backgrounds and philosophies. The primary purposes of the discussion are (1) to focus attention upon the issues and (2) to share some of the experiences and ideas that are taking place in the field.

In addition to examples of competency-based programs in the literature, reference will be made to some programs whose components were described or expanded upon by the practitioners involved in their implementation. A very brief description of some of the components now available in the literature follows.

The California State Department of Education administers the mandated California High School Proficiency Examination (CHSPE), which can now be taken by high school students who are second-semester sophomores or older. Any student who attains the proficiency level receives a certificate of proficiency which is legally equivalent to a high school diploma.

The test was developed by the California Department of Education and is administered and scored by ETS. The test is a functional literacy test containing many items from the Adult Performance Level Functional Literacy Test (APL), which was developed at the University of Texas at Austin for adult basic education; and it also contains other items testing language and math skills.

The proficiency level was set by testing a sample of high school seniors and setting the passing score at the 5[illegible] percentile. At the present time proficiency requires correctly answering approximately 75% of 150 multiple choice items, in addition to writing an acceptable essay.

The Michigan Educational Assessment Program is an on-going objectives-referenced assessment program which tests all students each year in reading and mathematics at grades four and seven. Student attainment of an objective is defined by correctly answering at least four of the five items, and results have been reported by objective for students, classrooms, schools and districts.

In 1976-77 Michigan added a new reporting scheme which is being commonly referred to as the "proportions report" (MDE, 1976a and b). This refers to the proportion of students (in the state, district or school) who attained an objective. At the state level value

judgments about objective attainment have been established as follows:

1. acceptable: 80-100% of the students attained an objective

2. encouraging: 60-79% attained an objective

3. low: 40-59% attained an objective

4. unacceptable: 0-39% attained an objective

These categories were established arbitrarily and apply to both reading and mathematics at grades four and seven. Across both grades and subject areas approximately 37% of the objectives were in the "acceptable" category, and the only area with objectives categorized as "unacceptable" was seventh-grade mathematics.

The Denver Public Schools have required for 16 years that high school graduates demonstrate proficiency in four separate areas -- language, reading, spelling and arithmetic -- which are represented by four sub-tests of a proficiency test. An overview of the Denver program was presented by Clark and Thomson (1976, pp. 7-8). The cut-off score on each sub-test is 32 of 50 items, and this was set arbitrarily after studying test results over a 3-year period. The program was initiated to insure that all graduates would have a good set of basic skills as they entered the job market. In a recent follow-up study of graduates, over 80% felt they had received good basic skills, but there is some pressure for raising the cut-off score to 35 of 50 items on each subtest.

The Palm Beach County Schools in Florida have mandated that each student must be functionally literate to receive a regular diploma. The test instrument is the APL which is now being further developed by ACT for publication.

This 75-item test of functional literacy is being normed with an adult population. Palm Beach County used the norming results, along with some results from supplemental studies conducted locally, to establish a cut-off point for proficiency. After judging that some groups of adults have less opportunity for leading productive lives (e.g., persons that are unemployed and on welfare, and persons in prisons), scores for adult population subgroups were examined. It was determined that a score above 53 could generally be associated with productive

adults and that a score of 53 or below could generally be associated with marginally productive adults.

Students will have an opportunity to take the proficiency test four times in their junior and senior years, and remedial programs will be available. Students who do not pass the test but fulfill other graduation requirements will receive a certificate of attendance, and they will have the opportunity to take the test after leaving school.

It is impossible to reach definitive conclusions from the small quantity of literature that is available about actual standard-setting procedures and interviews with selected practitioners; however, at this point in time there is value in utilizing informal qualitative research methods to see which problems and issues are of most concern to the practitioners and to help judge the status of standard-setting procedures in terms of some important considerations. Although this section of the paper is not organized according to Shepard's issues and recommendations, it must be noted that her identification of the potential problem areas served as an excellent point from which to study the standard-setting problem.

Before discussing issues and problems it is important to indicate the practitioners' environment. Those who are working with many of the competency-based programs are working under the pressures of time and politics. Since they are generally working in state or local education agencies, these practitioners almost always face the lack of resources associated with tight budgets; and they are usually caught up in the decision-making problems of bureaucracies. Their primary task is to work within these constraints and conduct a program that meets its major goals. One seldom finds a practitioner involved in implementing competency-based programs who is not aware of many of the measurement problems that surround competency-based test development and standard-setting, and who does not have some creative ideas about how to take them into account; but the practical constraints limit the degree to which this can be accomplished.

The Standard-Setting Role Within Competency-Based Programs

Spady and Mitchell (1977) made the generalization that substantial improvement in student outcomes from competency-based programs will not be the result of clarifying goals; rather it will derive from channeling the efforts of students and teachers for more effective learning. Clark and Thomson (1976) more specifically stated the following:

Since schools want proficient graduates, the administration of a competency test should be timed well in advance of graduation. When students do not reach the minimum level of competency, schools must respond with remedial programs to help the student reach these minimum levels. The payoff for competency testing is in appropriate remediation for students lacking essential competencies.

Competency tests should serve not only as an opportunity for students to identify deficiencies and to demonstrate important skills but, more importantly, to provide an impetus for revising program sequence and content to help students reach desired levels of proficiency. The purpose of competency tests is to "screen in" students, not to screen out students. (p.10)

Those who are involved in setting standards are generally sensitive to this perspective. They desire to make a positive contribution to the educational process, and they view the decisions of "passing" and "failing" as means rather than ends. In Denver approximately 40% of the ninth-grade students are below the proficiency level on at least one of the four subtests. Diagnostic profiles for these students are provided to teachers, and specific remedial programs are made available. By the senior year less than two percent fail to attain proficiency on all sub-tests, and many of these students do not meet other graduation requirements. Clark and Thomson cite other examples of remedial programs within competency-based programs at the local level.

A recent Colorado law requires that any district which imposes a special provision for graduation from the twelfth grade must (1) initially administer the test in the ninth grade, (2) offer the test at least twice during each school year, and (3) use the overall results of the tests to design regular or special classes to meet the needs of all children. Michigan plans to use the district "proportions" results to help determine which districts will receive state compensatory funds and what types of local programs will be developed.

At one school of Palm Beach County, which is just initiating its program, functional literacy test scores from the first year of testing were related to those on a standardized achievement test, and the probability of each junior failing the functional literacy test was calculated. All juniors with a probability of failure at 50% or higher were informed of this possibility, and special night courses were established for voluntary enrollment.

The example from Palm Beach County shows sensitivity to learner improvement and exhibits creativity, but it is also a response to one of the common problems with newly developed competency-based programs. In many cases political pressures have caused proficiency programs to be implemented immediately; therefore, tests and proficiency levels are developed under severe time constraints and initial remedial programs have some characteristics of "crash" programs. All of this may be worthwhile if the specific competency-based program becomes continuous and receives adequate resources for development and implementation; those juniors and seniors who have to undergo the added burden of catching-up to the new standards will probably be somewhat improved educationally, and most of the administrators and teachers will be more aware of some worthwhile theories and practices. However, if the program is a fad or is never adequately developed, the turmoil created in the district will almost certainly create more harm than good.

The Effects of Minimum Standards on the Limits of Educational Achievement

Spady (1977) warned that the tension between "required minimums" and "desirable maximums" is unavoidable in standard-setting, especially in cases where exit rules provide for full certification upon fulfillment of minimum standards. Shepard indicated that the theories and assumptions of competency-based education do not limit education attainment to the minimums, but that the popular interpretations of many legislators and citizens carry this possible implication.

There are two aspects to this problem: content and standards. In terms of content, if students are not required to be exposed to some areas (e.g., social studies, science, the arts) to graduate or if these areas receive much less instructional and evaluative emphasis, there is no way to assure reasonable attainment. In terms of standards, if proficiency levels are set at low levels and if exit with full legal rights is provided for those who meet these standards, it is possible that many students will choose this route; and it may be found at a later date that this has adversely affected the individual or society.

It is too early to determine the ultimate effects of this problem, but there is evidence worthy of notice. Most examples of competency-based programs are in the basic skills or functional literacy. Clark and Thomson listed 22 districts with competency-based programs, and all but seven were limited to these areas.

Unfortunately, the emphasis on basic skills is probably greater than it would appear on the surface. Even when higher level skills and non-basic content areas are included in program objectives, most educators who are aware of the realities of testing and measurement are suspicious that these may not be properly represented in the evaluation process.

Marlowe (1976) reviewed a copy of the California High School Proficiency Examination (CHSPE), and in an editorial-like article he criticized the test for being biased toward good readers. However, his major criticism was that the test did not include any items measuring a number of very important goals which are included in the goal-statements of nearly every local district.

The Task Force on Graduation Requirements of the National Association of Secondary School Principals (NASSP) has taken the position that the diploma should represent verified attributes in basic skills, acquaintance with the American experience, and completion of planned experiences including some in the group setting. They then state that some of these attributes can (and should) be verified by competency measures but that others, which cannot be verified by competency measures, should be verified by course credits and attendance records.

The Concept in Certification.

The need for statistical consideration of false positives and false negatives in specific competency-based programs depends upon the ultimate importance of the decision. Provision for early exit from high school must be considered ultimately more important than the diagnostic instructional situation, but there may be some protection from political and legal repercussions because a "reward" is being offered for passing rather than punishment being delivered for failing. The more potentially explosive situation occurs where students are prevented from receiving public high school diplomas because they have failed to achieve the passing score on a competency test. Based upon political and legal controversy related to the National Teachers Examination and other job certification tests, it can be predicted with near certainty that these results will be challenged by some persons who have failed to reach the proficiency level.

Another situation of great importance and potential consequence is

test discrimination against a population subgroup. In spite of the many conceptual and practical distinctions between norm-referenced and criterion-referenced testing, there is an abundance of evidence indicating that there will be great differences in the percents of students in various population subgroups who pass the various types of competency tests. Again, recent history indicates that these types of results will be severely challenged.

It seems highly probable that the most political and legal pressure will be applied by those representing students who claim to be "false negatives" or claim they are being discriminated against. In either case the pressure will be to lower the standard or by-pass it. If standards are used in mandated situations to make important decisions, and if the statistical considerations surrounding the reliability and validity of these standards fail to stand up under critical review, it is possible that these standards, which will already be considered minimal, will be lowered or by-passed. Under these conditions it is not only possible that valuable resources will be spent in legal battles and specific programs will be scuttled; it is also possible that a movement intended to assure that most students attain essential skills before leaving school will have the opposite effect of providing an escape route for many students who have acquired only a marginal education.

When pass-fail decisions lead to political and legal controversy, the standard-setting process will not be the only component of the specific competency-based program subjected to critical review in the political and legal decision-making processes. Other aspects of the program will be reviewed, and some of the aspects which will be of major concern include relevant program objectives, early detection and notification of weaknesses, timely and effective remediation programs, and consideration of test bias in instrument development.

The Gap Between Competency-Based Theory and Practice

There is a rather wide gap between standard-setting theory and practice in competency-based testing. Most practitioners involved in developing measurement procedures for competency-based programs lean decidedly toward the philosophies associated with criterion-referenced measurement; however, in important decision-making situations, practitioners have commonly turned toward norm-referenced procedures, setting relative standards that rely upon the performance of a criterion group. For example, California's CHSPE,

the only fully operable early-exit examination, uses the 50th percentile of a norming population as the performance standard; and the Palm Beach County Schools, where traditional norm-referenced interpretations were not acceptable, finally developed a standard-setting procedure which relies upon the mean scores of various criterion groups. In both of these cases the tests were constructed using procedures associated more with objectives-referenced philosophies than with norm-referenced philosophies. This gap is further represented by the fact that even when the performance standard has been set as an absolute, without concern for relative performance, it has almost always been established somewhat arbitrarily. By necessity, human judgment has been utilized in setting the standard; but the procedures for utilizing that human judgment have not usually been systematically developed and implemented; therefore, controversies have often arisen as to whether the absolute standards are fair and acceptable.

As long as performance standards have been established to make diagnostic instructional decisions within the educators' own area, practitioners have been able to successfully use absolute standards that are set rather arbitrarily. Controversies associated with test development and performance standards arise, but they can be resolved by educators. Also, in the interpretation guidelines – whether they be for a commercially developed test or a statewide assessment – the ultimate burden can, and will be placed on local administrators or teachers to re-interpret any questionable results by examining the test materials, setting their own standards, and making their own decisions.

There are many examples of the use of absolute standards in making various types of decisions at district and state levels. These examples even include the continued use of absolute standards in historically controversial situations where a limited amount of educational resources are distributed based upon results at the district and school levels, e.g., Michigan.

However, as the stakes have increased, practitioners have found that the necessary development and research have not been accomplished for providing criterion-referenced performance standards with which sound pass-fail decisions can be made and statistically defended. Some of the same concerns which cause only minor educational disagreements and points of controversy in instructional situations

suddenly cause severe problems when the decisions are more important and cannot be resolved within the educational arena. Some of the difficult questions which might arise from arbitrarily established absolute standards include:

1. If students must meet a statewide performance standard to graduate and 30% of the juniors fail, how is each school district to obtain and utilize the resources to successfully remediate most of them in one year? Will the performance standard be defensible in court? If 10% of the White juniors fail and 60% of the Black Juniors fail, will the test development and standard-setting procedures be defensible in terms of potential bias against Blacks? Will remediation components also be defensible?

2. If a statewide law mandates that students can leave school at age 16 upon successful completion of a proficiency test and it is found that 60% of the sophomores in the state can meet the criterion, is this acceptable? If 40% of the sophomores and juniors take the test, is this acceptable? If it is projected that 20% of the sophomores and juniors will take advantage of this avenue for early exit within the next five years, is this acceptable? Are the students who leave school via this program more likely to be "functionally literate" than they would have been under present requirements? What are the potential side effects of having these adolescents out of school?

These concerns are not conceptually peculiar to competency-based programs using absolute performance standards; but surprising results are much more likely to occur when arbitrary standards are set than when normative standards are used. The occurrence of potentially unacceptable consequences can be controlled by establishing some form of norm-referenced standard. Short run program effects can be made politically credible, and the additional resources needed for remediation can be predicted, through the utilization of normative procedures. Also, there has been some experience in validating norm-referenced standards; therefore, challenges to the standards can be defended more successfully.

Toward Closing the Gap: Consideration of Item Difficulty. If absolute performance standards are to provide more realistic, credible results and if they are to be more defensible as important (or sole) criteria in competency-based programs involving important decisions, one of two things has to happen. Either domain-referenced tests are going to have to become a practical reality, or those practitioners and measurement specialists involved in setting standards are going to have to develop empirically proven procedures that take item difficulty into account.

At the present time the most rigorous models concerning test length and estimation of true scores are based on the assumptions surrounding domain-referenced testing; and setting the passing score may be more manageable in domain-referenced testing because of the detailed specifications and procedures used to generate and select items. Unfortunately, as Millman (1974) acknowledged, there has not yet been enough experience in constructing domain-referenced tests to know if they can be successfully developed to provide criterion-referenced interpretation of many important educational achievements. It is also possible that domain-referenced testing will be more appropriate for diagnostic instructional purposes than for broader purposes such as program evaluation and decisions involving essential competencies. It would be very difficult to fulfill the demanding specifications of domain-referenced testing and also overcome some of the major practical problems recognized by Kosecoff and Fink (1976), e.g., too many narrowly defined objectives, too much testing time, and high costs.

It is certainly true that very few of the criterion-referenced tests currently being used fulfill the rather restrictive assumptions of domain-referenced testing. However, as noted earlier, passing scores are usually set either arbitrarily or with the overriding concern being test length and the "reasonable" probability of proper classification, accepting some of the unfulfilled assumptions of domain-referenced testing.

Many authors have discussed the necessity of taking item difficulty into account in criterion-referenced interpretations (Airasian & Madaus, 1972; Klein & Kosecoff, 1973; Millman, 1974; Quirk, 1974). Given the pervasive effects of item difficulty on test scores, those who set absolute criterion standards for objectives-referenced

tests must deal directly with the items used to measure objectives; and they must be provided some form of normative information relating to item difficulty.

Although there have been various attempts to develop and implement *a priori* standard-setting models which take item difficulty into account, no generally accepted model has emerged for use in competency-based programs. Some models have attempted to account for item difficulty in conjunction with item importance or face validity, while others have separated these concerns. Two models which deal with item difficulty separately are discussed briefly.

One such model has been used extensively in medical education. The method was developed by Nedelsky (1954) and involves teachers in judging the distractors which minimally passing students will be able to identify as incorrect. Meskauskas (1976) reviewed several articles concerning the application of this model, and Smilansky and Guerin (1976) recently reported promising information about the validity of the *a priori* judgmental process as related to empirical item difficulty and to the choice of distractors by students of varying achievement levels. Meskauskas concluded that this method has usefulness but that there must be a broad base of consensus to insure that a common standard is established.

Weiss and Conaway (1976) discussed procedures for determining desired performance levels and minimal acceptable levels of performance across objectives on an objectives-referenced test. This procedure involved teachers in predicting the percent-correct for each item and setting desired and minimal levels for each item. These values are then averaged across items and compared with actual results, by objective. The procedures, which were developed for use in educational assessment programs, have been applied to group performance results in several state and local assessment programs to determine instructional strengths and weaknesses. The procedures have been adapted to provide (1) consideration of the *a priori* standards along with independent *ex post facto* judgments based on actual item results, and (2) interpretation of the judgmental results along with normative results. Weiss and Conaway have also made available some information relating to agreement of teacher predictions with actual item results and comparing the consensus judgments of different groups of teachers for the same items.

These works do not provide practitioners with fully validated models for setting standards, but they do indicate that it is possible to develop systematic consensus procedures which will take item difficulty into account. If practitioners are to develop criterion-referenced standards that will have any chance of being both realistic and statistically defensible, they will have to conduct empirical and statistical studies in two areas: human judgment in setting standards, as it relates to item difficulty; and the use of normative data in refining these group judgments.

Chapter V

Instrument Quality Indices in Competency-Based Programs

The issue of whether the classical measures of validity and reliability are directly relevant to competency-based tests has been debated. A prominent feature of the debate centers around the score on the competency-based instrument and the use made of that score. The papers which follow in this section have both been authored by Robert L. Linn. The validity paper represents a combination of his papers from two sessions at the NCME Annual Meetings which were sponsored by the Task Force and a paper written at the Center for the Study of Evaluation (UCLA) while Dr. Linn was a visiting scholar supported by funds of the National Institute for Education. The reliability paper is based primarily on his work at the Center.

We believe that the papers in this section not only trace the history of the problems of identifying an appropriate index of reliability and validity, but that they also point to directions for solutions to the complex problem of gathering evidence to support the quality of a measure. Dr. Linn puts the problem of technical quality into the context of the measurement literature which has to do with tests which are designed to make a dichotomous decision and which are not necessarily designed to produce a great deal of variance. In this context, competency-based tests are much like criterion-referenced tests and objective-refererenced tests.

Issues of Reliability in Measurement For Competency-Based Programs

Robert L. Linn
University of Illinois, Urbana-Champaign

During the past decade there has been some controversy about the procedures for obtaining reliability estimates of criterion-referenced measures (here read competency tests). Part of the controversy has stemmed from questions concerning the role of score variability in traditional procedures for estimating reliability versus its role in criterion-referenced measurement. The use, if any, to be made of a criterion score in the estimation of reliability should be based on the metric of the measure or on the consistency of the categorizations resulting from use of the measure. Other issues could, of course, be identified, but the other issues can generally be subsumed under, or be viewed as closely related to, one of the above.

Score Variability

As is true with many aspects of competency measurement, the roots of the score variability issue can be traced back to discussions in the measurement literature that appeared before competency measurement became a major focus of attention. Lindquist (1936), for example, distinguished between tests designed to "rank the students tested" and those designed to "discover specific weaknesses, errors, or gaps in the student's achievement" (p. 20). For the former purpose, score variability would be sought and only discriminating items" answered by some but not all pupils in the group being tested" (p. 30) would be included in the test. For the diagnostic test, which was described in terms largely in tune with today's descriptions of criterion-referenced measures, however, it was seen as "legitimate to include any items about which information concerning the achievement of pupils is desired" (p. 30), whether or not they were answered correctly or incorrectly by all students in the group to be tested. The result might be a test with little variability which would appear in an unfavorable light if traditional reliability estimates were used as the standard for judging the test.

Other early discussions related to the issue of score variability for criterion-referenced measures could be noted, but it wasn't until 1969 that Popham and Husek focused attention on score variability

specifically within the context of criterion-referenced measurement. They argued that "the issue of score variability is at the core of the difference between norm-referenced and criterion-referenced tests" (1969, p. 3). Variability is essential in order to discriminate among individuals and is something that is sought after in norm-referenced measures. For criterion-referenced measures, however, score variability is seen neither as desirable nor undesirable; it is simply irrelevant. "It is, of course, true that one almost always gets variant scores on any psychological test; but that variability is not a necesary condition for a good criterion-referenced test" (p. 3).

Although variability may not be a necessary condition for a good competency measure, it is a necessary condition for obtaining high traditional reliability estimates whether they be of the internal consistency, parallel forms, or test-retest variety. The dependence of correlations on variability is well known; reduced variability can generally be expected to yield lower correlations. Thus, reliability estimates based on the correlation of parallel forms or on a test-retest correlation are viewed as suspect for competency measures. High correlations would be fine, but are not necessary, for good competency measures. In the presence of a low correlation, one would not know whether this was due to a faulty measure or simply to the lack of variability of a perfectly acceptable measure. Similar comments could be made about traditional measures of internal consistency, since they too depend on variability.

For Popham and Husek the dependence of traditional reliability measures on score variability caused a serious dilemma because internal consistency and stability were viewed as virtues for a criterion-referenced measure, but no clear way of demonstrating the lack of these virtues was seen in the presence of limited-score variability. Their position and acknowledgement of the dilemma with regard to internal consistency was clearly stated:

> Now it is obvious that a criterion-referenced test should be internally consistent. If we argue that the items are tied to a criterion, then certainly the items should be quite similar in terms of what they are measuring. But, although it is obvious that a criterion-referenced test should be internally consistent, it is not obvious how to assess internal consistency. The classical procedures are not appropriate. This is true because they are dependent on score variability. (p. 5)

While clearly stating the dilemma, Popham and Husek did not provide guidance on how to deal with it other than a half-hearted suggestion that confidence intervals for individual scores might provide a "partial solution." As noted by Simon (1969), however, even the suggested confidence intervals would be dependent upon some type of variability, either within or between persons.

To some, the issue of limited-score variability has been considered to be of little, if any, practical importance. The divergence of opinions was foreshadowed in an earlier debate between Horn (1966, 1968) and Ebel (1968) in the context of tests designed for purposes of assessment rather than prediction. As was later argued for competency measures, Horn claimed that the overriding concern for an assessment measure is representativeness of content and the scores on such a measure need not have variance. Ebel, on the other hand, argued that an assessment measure should have variance. He then went on to try to dismiss the issue by noting that "of course any debate over whether assessment devices **should** yield variant scores is slightly academic because in practice they almost always **did** yield variant scores" (p. 71, emphasis in the original). Little harm to the positions would result from substituting "competency measures" for "assessment devices" in the above exchange.

In a more recent exchange, Woodson (1974a, 1974b) suggested that while an item and/or a test may lack variance for a particular sample (e.g., a group of all masters), both items and tests should have variance for ". . .the population of observations for which a test is designed." (1974a, p.63) (e.g., a group consisting of both those designated competent and those who were not). Millman and Popham (1974), however, responded by saying that "item and test score variability is an unnecessary requirement" (p. 137) for a criterion-referenced measure.

Both sides of the controversy readily acknowledge that score variability will almost always be found in practice. The crux of the difference lies in the importance attached to the notion that variability is not a necessary consequence of a good competency measure. If zero score variability may be expected to occur rarely in practice, then why has it received all of the attention it has? There probably are several answers to this question. In part the debate may be viewed as one of interest "in theory" if not in practice. But, there are more fundamental answers which make the debate an important one.

100

To begin with, a requirement of item and score variability may be seen as a threat to the fundamental principles for the construction of criterion-referenced measures. Clear definition of the content domain and representativeness of the sample of items from that domain are essential to competency measurement. If the requirement of item variance resulted in the elimination of selected items from a test, it would destroy the representativeness.

> The primary purpose of criterion-referenced [here the same as competency] testing is descriptive wherein the task is to describe the current status of an examinee. The question asked is how well has the examinee performed in relation to some domain of tasks of which the test items are only a sample. This is clearly the use under discussion in the Popham and Husek article. . ., in Glaser's (1963) classic description of a criterion-referenced test, and in the writings of Ebel (1962) on content standard tests, Osburn (1968) on universe defined tests, and Hively, et al. (1973) on domain-referenced tests. (Millman & Popham, 1974, p. 137)

Representativeness of the sample for the well-defined domain is the key issue and variability of responses to a sampled item is irrelevant unless used as a guide in the selection of items, in which case, it would be a menace.

Another part of the answer may lie in a belief supported by notions of mastery learning which is in close harmony with much of the thinking about criterion-referenced measures. In mastery learning the ideal result of instruction in a given content domain would be total mastery by all those instructed and therefore no score variability.

A cynic might note yet another reason. Criterion-referenced measures tend to be quite short compared to the typical norm-referenced measure. A consequence of brevity alone would be generally lower reliability coefficients. Hence, the need to diminish the importance of reliability as estimated by traditional means.

Whatever the reasons for the emphasis on score variability in considering reliability of criterion-referenced measures, it has been important not only to theoretical discussions but to the search for alternate procedures. It has also been a point of concern to the National Education Association. In *Guidelines and Cautions for Considering Criterion-Referenced Testing* published by the NEA (1975),

the dependence of traditional reliability estimates on score variability was noted. It was then argued that "using these techniques causes the tests to discriminate in the same way as do items in standardized tests" (p. 10).

The issue of score variability remains largely unresolved. Attempts to side step it have not met with general acceptance (e.g., see Livingston, 1972a whose approach is described below). It is our view that the issue is a conceptually important one and it should be made clear that variability should not be the guiding principle nor be allowed to distort representativeness. At the same time, however, the use of traditional indices need pose no real danger if not allowed to distort the content coverage. If traditional reliability indices are high, this is informative. If they are low, they are less informative, especially if the variability of the test scores or of a substantial proportion of the items (see Shoemaker, 1969) is small. As with any meausre, the importance of various types of reliability evidence will depend upon the use of and inferences based upon the scores (American Psychological Association, 1974). For some purposes, the consistency of categorizing people as masters or non-masters will be of primary concern (see below). For other purposes, concern will be with the limits of generalization and procedures stemming from the formulation by Cronbach, Gleser, Nanda and Rajaratnam (1972) (see below).

Deviations from a Criterion Score

Livingston (1972a) focused on the notion of a fixed standard or criterion score as the distinguishing characteristic of a criterion-referenced measure (as it is in competency measurement). He then used classical concepts of reliability, but with deviations taken around the criterion score rather than the mean, to derive "criterion-referenced reliability" coefficients. Various procedures from classical test theory based on deviations of scores, X, from a mean, μ_X, were simply replaced by counterparts using deviations from the criterion score, C_X. Livingston also showed relationships between his new formulas and their classical counterparts. Thus, for example, the classical definition of reliability as the ratio of the true score variance, σ_T^2, to the observed score variance, σ_X^2.

$$\rho^2_{TX} = \frac{\sigma^2_T}{\sigma^2_X} \qquad [1]$$

The criterion-referenced counterpart derived by Livingston is shown to be equal to:

$$K^2_{XT_X} = \frac{\sigma^2_T + (\mu_X - C_X)^2}{\sigma^2_X + (\mu_X - C_X)^2} \qquad [2]$$

The Livingston formulation was seen as a solution to the problem caused by low score variability. Stanley (1971) noted that:

> Criterion-referenced tests are meant to be used in situations in which there may be no variation among the true scores of the examinees; these measures are intended not to discriminate among persons, but to discriminate each person's score from a fixed criterion score. Therefore, a criterion-referenced test can give reliable information even though its classically defined reliability coefficient equals zero (p. 435).

Stanley then went on to indicate that Livingston's procedure provides the solution to this problem.

The apparent advantage of the Livingston approach can be seen for the special case suggested by Stanley by a simple comparison of equations [1] and [2]. When $\sigma^2_T = 0$, the classical reliability given in [1] is either zero (if $\sigma^2_X \neq 0$) or is undefined (if $\sigma^2_X = 0$). The result from equation [2], Livingston's criterion-referenced reliability, will be greater than zero even when $\sigma^2_T = 0$, as long as $\mu_X \neq C_X$. This apparent advantage of the criterion-referenced reliability must be weighed against several disadvantages that have been discussed in the literature.

Harris (1972) noted that the fact that Livingston's reliability coefficient is larger (provided $\mu_X \neq C_X$) than the classical reliability coefficient "does **not** imply that a more dependable determination of whether or not a true score falls below (or exceeds) a given criterion value" (p. 29 [emphasis in the original]). The dependability is a function of the standard error of measurement rather than the reliability and, as Harris showed, the standard error of measurement is the same under the Livingston or the classical formulation. Thus, the larger criterion-referenced reliability may be seen as potentially misleading, or, at least, less informative than the standard error of

measurement. (Harris also made another criticism which is not considered here. See Harris (1972) and Livingston's (1972b) reply.) Shavelson, Block and Ravitch (1972) showed that the magnitude of the Livingston coefficient can be altered radically as a function of the size of the difference between the mean and the criterion score. As the difference, $\mu_X - C_X$, increases in absolute magnitude so does the magnitude of the Livingston coefficient. Shavelson et al. viewed this dependence on the size of the $\mu_X - C_X$ difference as an important limitation. Livingston (1972c), on the other hand sees this dependence as quite consistent with the purpose of a criterion-referenced measure. Coupled with the often arbitrary choice of a criterion score (see Chapter IV of this monograph on setting standards), however, the dependence of the coefficient on its choice would at least seem to demand a better than typical justification for the choice of C_X or the reporting of coefficients for a range of criterion scores.

Possibly the most fundamental objection to the Livingston coefficient is that provided by Hambleton and Novick (1973). They argue that, where the purpose of a criterion-referenced test is to dichotomize examinees into two categories, e.g., master and nonmaster, then the concern is not with the size of deviation of a person's score from the criterion score but with whether the person is dependably categorized. In the terms of Hambleton and Novick, Livingston employs a "squared error loss function" whereas they would prefer a "threshold loss." In other words, consistency of categorization by two parallel measures or at two points in time should be evaluated in terms of an all or nothing standard (0 if inconsistent and 1 if consistent), as would be the case with a threshold loss function, rather than by the square of the deviation of a score from the criterion score.

Consistency of Categorization

The framework for much of the more recent work on reliability for criterion-referenced or competency measures was provided by Hambleton and Novick in the article that was just mentioned. They suggested that "the traditional concepts of reliability and validity could be replaced by a complete decision-theoretic formulation" (1973, p. 168). For situations where a test is used to dichotomize people into two categories, the metric of primary interest for the test

is a zero-one score where nonmasters receive a zero and masters a one. Questions of reliability can then be addressed in terms of the zero-one metric rather than in terms of the number-right score of some other metric. Since product-moment correlations have undesirable properties for zero-one variables, they suggested that "an alternative measure of reliability might simply be the proportion of times the same decision would be made with. . .two parallel instruments" (p. 168).

Adjusted Measure of Agreement

The basic idea suggested by Hambleton and Novick was developed and extended by Swaminathan, Hambleton and Algina (1974). Also, see Hambleton, Swaminathan, Algina & Coulson (1978) for a recent review. Specifically, they defined "reliability of a criterion-referenced test as the measure of agreement between the decisions made in repeated test administrations" (p. 264). Swaminathan et al. considered using the simple proportion of agreement between decisions on the two occasions but rejected this as an index because it "does not take into account the agreement that could be expected on the basis of chance alone" (p. 264). For example, if a test identified 90% of the people as masters at time 1 and 95% at time 2, then the joint application would be expected to identify 85.5% of the people as masters on both occasions and 0.5% of the people as nonmasters on both occasions, for an overall agreement rate of .86, assuming that the two testings were independent.

To adjust for chance agreement, Swaminathan et al., proposed that Cohen's (1960) coefficient kappa be used. Kappa is simply the proportion of agreement corrected for chance. It is given by

$$\kappa = \frac{p_o - p_c}{1 - p_c}, \qquad [3]$$

where p_c equals the agreement expected by chance and p_o is the observed agreement. In the above example, if 89% of the people were identified as masters on both occasions and 4% identified as nonmasters on both occasions, the p_o should equal .93. As previously indicated, p_c would be .86. The resulting Kappa would be .50. The simple

proportion of agreement of .93 might yield a rather different impression than the "corrected for chance" proportion of agreement of about .5.

Coefficient kappa is conceptually appealing as an index for a measure that is used to categorize people. It can be applied with two or more categories. With more than two categories, the seriousness of the disagreements can be differentially valued by using weighted kappa (Cohen, 1968). Kappa can range from -1 to +1 with a positive one indicating perfect consistency and zero indicating chance agreement. A value of one is possible only when the marginals are equal. Negative values imply that the observed agreement is worse that would be expected on the basis of chance. Sampling variances of kappa and weighted kappa have been derived (Everitt, 1968; Fleiss, Cohen, & Everitt, 1969) and Light (1973) has proposed a large sample significance test. The relationship between weighted kappa and the intraclass correlation is shown by Fleiss and Cohen (1973).

When the two categorizations are independent in the population, the expected value of kappa is zero. In the non-null case, however, kappa is a biased estimator (Harris & Pearlman, 1977). For this and other reasons, Harris and Pearlman prefer the unadjusted proportion of agreement.

Whether kappa or the unadjusted proportion of agreement is used, the approach to reliability suggested by Swaminathan et al. (1974) requires dual measurement, either test-retest or parallel forms. For many practical situations this is a severe limitation because it is often feasible to administer only a single test at one point in time. Consequently, there have been several efforts to extend the Swaminathan et al. approach to a single test administration.

Single Administration Reliability

Marshall and Haertel (1975; also see Marshall, 1975) proposed a single administration reliability based on the proportion of agreements which they called coefficient beta. Beta is the mean of all possible split-half coefficients of agreement. Thus, it could be obtained by dividing the test into all possible pairs of half-tests. The categorization of examinees on the basis of each half-test in a pair would be compared and the simple proportion of all classifications that agreed would be computed. These proportions would then be averaged over

all split-halfs. To estimate the reliability of the full length test they suggested that beta simply be substituted into an empirically developed formula that was found to approximate the mean of the values obtained using the Spearman-Brown formula. Analogy is apparently the only justification for substitution into the Spearman-Brown formula in the first place and thus the estimate for the full length test is left in doubt.

More recently, two single administration procedures for estimating reliability have been suggested (Huynh, 1976; Subkoviak, 1976). Both approaches are based on the idea first put forth by Hambleton and Novick (1973) that the reliability of a decision-oriented measure is best judged in terms of consistency of the decisions. Thus, they are similar in philosophy to Swaminathan et al. (1974), Huynh uses kappa, while Subkoviak joins Marshall and Haertel in estimating an unadjusted proportion of agreement.

Huynh's approach is derived from a model in which the probability of a given number of items correct for an examinee is assumed to be given by a binomial density function with a parameter equal to the examinee's true score, i.e., the proportion of items in the complete domain that would be answered correctly by the examinee if all items in the domain could be administered. This assumption amounts to assuming that all items in the domain are "exchangeable." In other words, it makes no difference which subset of items are selected for an examinee. The true scores are assumed to follow a beta distribution. Together these assumptions yield a "beta-binomial" model which has proved useful in applications of Bayesian statistics (Novick & Jackson, 1974). For large samples, Huynh suggested that the parameters of the beta distribution of true scores be estimated from statistics of the observed test score distribution (see his equations 2 and 3). The joint distribution of the test and a hypothetical parallel form can then be estimated using the beta-binomial model. This estimated joint distribution can then be used to estimate the proportion of agreement, which in combination with model estimates of the marginal distribution can be used to estimate kappa.

The model proposed by Huynh is a relatively strong one and requires heavy calculations. The heavy caluculations and complexity may discourage many potential users. He did provide an approximation procedure when a moderately large number of items (say 10 or more)

are available, which should extend its utility. Even 10 items are more than is available on many criterion-refererenced tests, but this is often a low estimate of the number of items on a competency test.

The assumption that the items are exchangeable may be questionable. As suggested by Lord and Novick (1968) the compound binomial model may be better than the simple binomial model. But, "for many purposes, the binomial error model can be viewed as a useful approximation" (p. 524).

The magnitude of kappa is shown to depend on the criterion score used to form categories, on test length, and on score variability (Huynh, 1976, pp 260-261). Kappa tends to be lower for criterion scores near the extremes, to increase with test length, and to increase with test score variability. Since kappa is not unique for a test but depends on the cut-score that is used to form categories, Millman (1974) recommended that description of the specific situation, including the cut-score used, be included in reporting kappa.

The approach suggested by Subkoviak (1976) has some close similarities to the Huynh approach. He too was concerned with consistency of decisions and assumed a binomial model for the observed scores of each examinee based on the examinee's underlying true score. Subkoviak proposed that proportion of agreement be used rather than kappa, however. He also suggested some simpler approaches to estimating the true proportion of items correct than the beta distribution approach used by Huynh. The true proportion correct would be estimated either by the observed proportion of items correct for each individual, which is the maximum likelihood estimate, or by a regression estimate based on the observed sample mean and the sample Kuder-Richardson Formula 21 reliability coefficient. They estimate probabilities that a person's score is above or below a given criterion score for the actual test under consideration and for a hypothetical parallel test. Together these estimated probabilities are used to estimate the proportion of agreement.

Subkoviak also has developed procedures based on the compound binomial model in place of the simple binomial model. He compared the estimates using the binomial model and the compound binomial model for a single test administration with estimates based on a dual administration. Both single administration procedures slightly underestimated the dual administration estimates at all cut scores. The

compound binomial model was somewhat better than the simple binomial model, but the latter performed quite well and would seem to provide a useful approximation.

The effect of using three different estimates of true proportion correct scores on Subkoviak's proportion of agreement estimates were compared by Noe and Algina (1977). Based on their simulation results, they concluded that regression estimates based on KR-20 provided generally accurate estimates with reasonably small sampling variability.

As was noted above, the magnitude of kappa is dependent on the choice of a cut score. The same is true of the uncorrected proportion of agreement. Subkoviak obtained higher proportions of agreement for all three estimation procedures that he investigated with cut scores near the extremes than with intermediate values. In view of the dependence of the coefficients on the choice of criterion score, Huynh (1977) proposed that the maximum value of kappa be estimated over the range of possible criterion scores. The resulting coefficient is called kappamax. He also provided formulas of the standard error of kappamax and approximate formulas for the kappamax of a double-length test. Although kappamax has the advantage that it would take on only one value for the sample results for a single test rather than different values for each possible criterion score, it does not correspond to any specific decision rule. Consequently, kappamax may be high while the kappa corresponding to a particular decision rule was relatively low. Thus, kappa coupled with a description of the associated decision rule, as was recommended by Millman (1974), has greater practical utility.

Moyer (1976, 1977) conducted a study comparing reliability results based on traditional procedures (Kuder-Richardson Formula 20) and those based on consistency of classification (kappa). She found that the two approaches yielded very similar information for the tests used in the Michigan Assessment. Similar results were reported by Strasler and Raeth (1977). The tests that had relatively high (low) KR-20's also had relatively high (low) kappas. Moyer (1976) concluded that for practical purposes KR-20 was preferable since it is more familiar and yielded results similar to kappa. Since the two are conceptually quite different, however, this empirical outcome does not provide a persuasive reason for using KR-20 in place of kappa.

The choice should be dictated by the purpose (decision consistency or internal consistency) rather than familiarity despite the similarity of results in specific situations.

The Swaminathan et al. (1974) paper, together with those by Huynh and Subkoviak, provide the means of estimating the degree of decision consistency in either dual or single administration situations. They represent what seems to be a growing consensus that decision consistency is the main reliability issue for criterion-referenced measures or competency measures. As recently argued by Livingston (1976), however, for some purposes there may be an interest in something besides the reliability of classification. If so, Livingston argues that the coefficient he previously developed (1972a) provides an alternative where the focus is on the "reliability of measurement" rather than the "reliability of classification."

Additional Alternatives in the Literature

Edumetric Reliability

In addition to the proposals made by Livingston (1972a) and those based upon notions of decision consistency, a few other approaches to questions of reliability have been suggested within a criterion-referenced context. Carver (1974) suggested that the reliability of a criterion-referenced measure could be assessed by administering the test to "two comparable groups" (p. 56) and comparing the percentages of students in the two groups that exceeded a criterion-score. Alternatively, percentages of a single group exceeding a criterion-score could be compared (Carver, 1970). These suggestions fail to take into account the extent to which people are consistently classified. The proportion of people in a given category could be the same based on alternate forms without agreement as to who belongs to a given category. Thus, as noted by Hambleton et al. (1978) Carver's "conditions are necessary but not sufficient to establish test reliability"(p. 21). The procedures for getting at decision consistency (e.g., kappa or proportion agreement) are to be preferred.

Carver (1974) also made some suggestions for what he called the "edumetric dimension" of tests. Although distinct from concepts of competency measurement, his edumetrics would appear to have relevance in the competency context because he was concerned with absolute rather than comparative decisions. He also claimed that his "edumetric reliability is not dependent on variances of test scores as in psychometric reliability" (p. 514).

According to Carver (1974), reliability "is conceptually a synonym for consistency" (p. 514). Unlike classical psychometric procedures where there is a concern for consistency of discrimination among persons, Carver suggested that for edumetric purposes the concern is with the consistency of within individual change. He proposed that edumetric reliability be evaluated by comparing the consistency of gains from pretest to posttest on parallel forms administered at both points in time. Alternatively, he suggested that parallel forms be administered "under equal treatment conditions. . .[in order to determine] the degree to which individual scores are the same on both forms" (p. 514). Haladyna (1975) has argued, however, that "standard errors of measurement are not reliant upon the variability of test scores" (p. 603). Furthermore, as noted by Cronbach (1975), estimated components of variance are useful not only for comparative decisions but for absolute decisions.

Coefficient of Reproducibility

Cox and Graham (1966) suggested the coefficient of reproducibility as an alternative to traditional procedures for estimating the reliability of a decision-oriented measure. The coefficient of reproducibility is appropriate for items that are sequentially scaled such that a person is expected to pass items up to a certain point and fail items beyond that point. In other words, the items should be expected to approximate a Guttman scale (see Torgerson, 1958, Chapter 12). This would not seem to be a reasonable expectation for most competency measures. Certainly, it is far removed from the assumption of exchangeability of items used by Huynh (1976). For more on reproducibility and criterion-referenced measurement, see Hambleton and Novick (1973, p. 167).

Index of Efficiency

The final approach to be mentioned in this section is the index of efficiency suggested by Harris (1974). The index of efficiency is relevant to notions of reliability but was never intended "to mean all that reliability means" (p. 106). Since criterion-referenced measures are often used to make dichotomous mastery decisions (i.e., to serve as competency tests) Harris reasoned that a coefficient that indicated the extent to which a test discriminates between masters and nonmasters would be informative. The index he developed is defined as the ratio of the between-group (i.e., masters vs. nonmasters) sum of squares to the total group sum of squares. The check on the test's ability to discriminate between categories provided by the index of

efficiency should be useful. For many purposes, however, the squared deviations used in the index may not be the right metric. Two non-overlapping distributions of scores for masters and nonmasters could still have substantial within-group variability and therefore an index of efficiency substantially less than one, even though members of the two groups could be discriminated with perfect accuracy.

Generalizability Theory

Generalizability theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972) provides a framework for evaluating the dependability of measures that is relevant not only in a norm-referenced context but in a criterion-referenced one as well (Brennan & Kane, 1977; Hambleton et al., 1978). It deserves more attention for addressing reliability questions concerning competency measures than it has received to date. In generalizability theory the focus is appropriately on **components of variance** rather than on coefficients of generalizability. The focus on components of variability may seem contradictory to the rationale of criterion-measures where, as was discussed at length above, score variability is considered irrelevant. Certainly if there is no score variability then all components of variance must also be zero. What use, then, does generalizability theory have for criterion-referenced measurement? In short, the simple answer is "Plenty."

To see the relevance of generalizability theory, it is useful to recall the four types of interpretations based on the results of a measure that are considered by Cronback et al. Measures may be used (1) to make "absolute decisions," (2) to make a "comparison between two courses of action for individual," (3) to make a "comparison between persons" and (4) to draw a "conclusion about the relation between pairs of variables" (pp. 14-15). Only the third type of interpretation falls clearly within the stereotypic mold of a norm-referenced procedure. The first is much more similar to descriptions of competency measurement than to norm-referenced measurement.

Cronbach et al. described an absolute decision as follows:

> Where an individual is to be classified in some way, performance standards may be set which determine how he will be treated. For example, a test is given an applicant for a driver's license, with the predetermined rule that a score of 85% is considered adequate for licensing. In making the decision, this person is **considered by himself** (p. 14, emphasis added).

All that is missing for the devotee of criterion-referenced measurement is the careful description of the item domain and procedures for ensuring representativeness. Adding these would do nothing to change the absolute character of the decision in contrast to a comparative decision. Neither would the addition alter the relevance of the statistical procedures developed by Cronbach et al. for application in this context.

The relevance of generalizability theory to criterion-referenced measurement was clearly recognized by Cronbach et al. They noted that:

> Almost all the uses of tests that are now attracting increasing interest demand absolute content-referenced, or criterion-referenced measurement. It is the level of the graduates' performance, and not their standings relative to each other, that one requires to evaluate early childhood education. It is absolute measurement that one uses to prescribe next week's instructional activities in a computer-based instruction, and interesting to see if a standard is met. . . .The emphasis in generalizability theory is on the universe score itself and not on comparisons of individuals stressed by classical test theory. (p. 384)

Thus, generalizability theory is purported to be of value for criterion-referenced measurement. Indeed there is a natural affinity between the criterion-referenced emphasis on domain definition, item sampling for representativeness and making inferences from performance on sampled items about mastery of a domain; and their counterparts with different labels in generalizability theory. Thus, one might refer to the universe of admissible observation, sampling from that universe, and making inferences about a universe score.

The array of estimation procedures encountered in generalizability theory is too great to review here. Basic to all of them, however, is the estimation of components of variance. Components of variance are informative regardless of the amount of score variability. For example, components of variance due to item form, time of testing or conditions of test administration are all potentially pertinent for a competency measure. If item form had a large component of variance associated with it, then one might consider limiting a domain (and thereby the generalization) to a single item form or using item form to stratify for item sampling purposes in order to enhance representativeness.

The relevance of generalizability theory for decision-oriented measures was recognized in a recent article by Brennan and Kane (1977).

The proposal in their article was not simply for an application of generalizability theory, however, but for an adaptation of the theory in a manner parallel to the adaptation of classical reliability estimates that was suggested by Livingston (1972a). Thus, Brennan and Kane argued that the primary interest for a measure is "in the difference between each examinee's universe score and the absolute standard defined by the mastery cutting score" (1977, p. 283). Consistent with this interest, Brennan and Kane proposed deviating observed and universe scores from the standard rather than from the mean.

The Brennan and Kane approach results in an "index of dependability" that differs from the usual generalizability coefficient as a function of the difference between the cutting score and the mean. Their dependability index is less than the usual generalizability coefficient when the cutting score is near the mean but greater than the usual generalizability coefficient for cutting scores far from the mean.

In generalizability theory, it is the estimate of components of variance rather than generalizability coefficients that are of primary interest. The estimate of the components of variance are not altered by the introduction of a cutting score for purposes of estimating an index of dependability. Thus, the potential utility of generalizability theory does not depend on the acceptance of a cutting score or on a coefficient based on deviations from it.

It is unfortunate that the issue of score variability for criterion-referenced measures has apparently distracted attention from the uses of generalizability theory for such measures. "Estimates of components are **particularly** illuminating when an instrument is used for absolute measurement" (Cronbach, 1975, p. 602).

Summary

In this paper on reliability I have discussed the issue of score variability for a competency measure. While concurring with the position that variability is not a necessary requirement for a good decision-oriented measure, I think that this position has mistakenly been interpreted to imply that certain approaches to assessing the dependability of a measure are inapproriate when the approaches are actually quite appropriate. Generalizability theory is the primary case in

point. It has great relevance for competency measurement but has received relatively little attention in this context.

I also reviewed a variety of procedures that have been specifically suggested for assessing reliability of criterion-referenced measures. Of these procedures the ones that have built upon the idea of assessing the consistency of decisions (Hambleton & Novick, 1973) appear to have the greatest utility. They are most useful in contexts where a measure is used to categorize people in two or more (but a relatively small number) of groups. Notable in this connection are the papers by Swaminathan, Hambleton, and Algina (1974), Huynh (1976) and Subkoviak (1976).

Issues of Validity in Measurement For Competency-Based Programs

Robert L. Linn
University of Illinois, Urbana-Champaign

At the 1976 annual meeting, I participated in a debate organized by the NCME Task Force on Measurement Problems in Competency-Based Programs. I argued then that the primary measurement issues in competency-based education "are not fundamentally different than those encountered in other contexts" (Linn, 1976). The primary issues in competency measurement are validity and generalizability. These are also the key issues in measurement of human behavior in contexts outside of competency-based programs.

Although I think now as I did then that the primary measurement issues are the same in competency-based programs as in other contexts, these issues become more salient and urgent in competency-based programs than in traditional programs. It might also be noted that the measurement problems appear more overwhelming in the competency-based context. The greater urgency and awesomeness of the measurement problems are due to the fact that measurement plays a much more central role in these programs than in traditional programs. Indeed, measurement is at the heart of such programs.

Another distinguishing characteristic of the measurement requirements of competency-based programs is the emphasis on performance. Old faithful, the multiple-choice question, just isn't adequate for many of the competencies of interest. Performance measures may need to involve operant behavior rather than just respondent behavior (McClelland, 1973), and will often be ill-suited for the paper-and-pencil format. Switching from paper-and-pencil tests to observations of performance or rating of products doesn't change the fundamental issues of validity and generalizability but it does place more difficult demands on the measurement specialist.

In addition to placing greater reliance on measurement and making more difficult demands, some of the specific requirements are also different than those with which the traditionalist is used to dealing. The setting of standards and the reduction of the measures to dichotomies, for example, raise important issues (See Chapter IV of this

monograph) and require special analytical techniques. I'll make a few comments about the issue of standards and treatment of the resulting dichotomies but only as they relate to my central topic, validity.

The validity of a measure is often defined as the accuracy with which an instrument measures what it purports to measure. A broader and more useful conception is that questions of validity are questions for the soundness of the interpretations of a measure (American Psychological Association, 1974; Cronbach, 1971). Thus, it is the interpretation rather than the measure that is validated. Measurement results may have many interpretations which differ in their degree of validity and in the type of evidence required for the validation process.

There are a variety of types of validity questions which conventionally are addressed by different types of validation, e.g., criterion-related, content and construct. As noted in the *Standards for Educational and Psychological Tests*, "these aspects of validity can be discussed independently, but only for convenience. They are interrelated operationally and logically; only rarely is one of them alone important in a particular situation" (APA, 1974, p. 26).

The Variety of Validation Needs for Criterion-Referenced Measures

Millman (1974) described validity in terms of the "accuracy of inferences made from test scores" (p. 360). This position is generally consonent with the one stated above. While recognizing a variety of types of inferences requiring different validation processes, Millman argued that the most appropriate inference for a domain-referenced test concerns the status of an individual relative to a well-defined domain. For this inference, the logical analysis of the domain definition and procedures for generating items are of primary concern. That is, content validation is the primary issue.

I do not quarrel with the emphasis placed on content validity by Millman. Indeed, I think that one of the main advantages offered by competency measurement is that it provides a much firmer basis for content validation than has been available with traditional psychometric approaches to constructing achievement tests.

Content validity is a necessary consideration for criterion-referenced measures. It is my contention, however, that it is rarely sufficient. This is so because rarely, if ever, are inferences restricted to the type

that are adequately supported by content validity alone.

The use of a measure to classify someone as competent or incompetent requires that several questions of validity be addressed. Content validation is needed to insure that the set of tasks or observations actually included in the measure are an adequate representation of the domain of interest. Content validity depends heavily on the precision with which the domain is defined. Precise definitions are also crucial to making judgments regarding the importance of performance that is being measured and whether important aspects of performance are excluded. Content validity is very important for measurement in competency-based programs, but it is not sufficient.

With a well- defined domain and random sampling of items from that domain there is a firm basis for supporting an inference that a person can satisfactorily perform at least x percent of the tasks in that domain. This is the type of inference that is supported by content validity. It is a natural consequence of the adequacy of the domain definition and of item generation rules. Seldom are our inferences limited to such a degree, however. Labels such as "master" or " competent" are often attached to persons whose performance is estimated to be above some specified level. Scores on the test are apt to be used as the basis of differential treatment. For example, some students may be given remedial instruction as a result of their scores. These additional interpretations and uses require more than content validity (Linn, 1976).

When a person is labeled a "master" certain expectations usually accompany that label. The student who has mastered certain types of addition problems may be expected to be "ready" to learn subtraction. Mastery of one skill may be expected to be prerequisite for mastering another skill. Evidence of the soundness of expectations such as these is needed. The process of marshalling the evidence for the soundness of the expectations is construct validation (Cronbach, 1971; Messick, 1975).

Even where constructs are not invoked more than content validity will usually be needed. The decision to have students with scores on a test below some specified level repeat an instructional segment rather than go on to a new one is based on an implicit prediction that the student will, at least, learn the content of the segment to be repeated better by redoing it than by moving on. It may also be based

on an implicit prediction that he or she would do more poorly on the next segment than would otherwise be the case. Evidence of criterion-related validity is needed to support these predictions.

I have argued that all three of the traditionally recognized types of validity are relevant to competency measurement just as they are to other types of measurement. It does not follow from this position, however, that the traditional validation procedures will be the most appropriate ones for criterion-referenced measures. For instance, there are a variety of reasons that the traditional reliance on correlation coefficients may be inappropriate for competency measures. One of these is the issue of variability which is discussed at length in the reliability section of the monograph. Another reason is that, as suggested by Hambleton and Novick (1973), correlations are based on an inappropriate loss function for situations commonly encountered with criterion-referenced measures. Hambleton and Novick suggested that a threshold loss is more appropriate in a criterion-referenced context than is the traditional squared-error loss.

The need for a different type of statistics or for nontraditional procedures for collecting the needed data do not require different concepts of validation. They merely represent different details of implementation within a common conceptual framework. A few comments regarding the special implementation requirements for validation of criterion-referenced measurement may be useful, however. These are offered below under the three traditional headings of content, criterion-related, and construct validity.

Content Validity

Acknowledging the need for multiple types of validity information, let me return to content validity in the hope of becoming a little more specific. The key to content validation is the definition of the domain of interest. Ideally, the definition should include a description of every detail of the measurement procedure that might have an influence on the results. The ideal can never be fully achieved, but we could certainly improve upon current practice. Consider, for example, the observations of student teachers in a classroom setting. There is an almost limitless list of potentially important details that might have an influence on the results of the observation procedures. There are the physical conditions of the situation (e.g., building, lighting, room size, furniture arrangement, etc.). There are consider-

ations of timing (e.g., time of day, day of week, part of the semester). There are the students and other persons in the room to be considered (e.g., number, age, sex, etc.). The topic of the lesson, the amount of previous contact of the student teacher with the students, the obtrusiveness of the observers, and the characteristics of the observers all require attention. The list could go on and on. The description can never be complete. Judgment is always required in determining what is most deserving of attention, i.e., those aspects of the measurement procedure that are most crucial to the results of the measurement.

Certain aspects of the procedure may be controlled. For example, in the observation of student teachers, the content of the lesson, the number of students and the length of time available for the lesson, among a number of other conditions, might be fixed for everyone. All the possible variations of other conditions not fixed by the definition, e.g., observers, physical location, etc., constitute what Cronbach et al. (1972) refer to as "the universe of admissible observations."

It is possibly more common in discussions of content validity to speak in terms of a "performance domain" (e.g., APA, 1974) rather than a universe of admissible observations. The latter label, though longer and more clumsy to use all the time, has the advantage of calling attention to the idea that any characteristic of the measurement procedure which is not controlled in the definition, could be varied, and the observation would still be admissible. For example, the domain of all addition problems involving two single digit positive numbers appears fairly clear and specific. Such a definition, however, would include as admissible observations oral as well as written presentations, base eight as well as base ten numbers, speeded as well as unspeeded presentations, multiple-choice as well as constructed response formats, and many other variations. The universe of admissible observations calls attention to the limits and possible vagaries of the definition of the measurement procedure that may go unattended in the description of a performance domain.

Definition of the universe of admissible observations is the first of two crucial steps in content validation. The second step requires a demonstration that the observations yield results that are representative of that universe. Where the universe is sufficiently well defined, representativeness may be pursued by a formal sampling

process. Random or stratified random sampling has great advantages of objectivity and enables one to calculate unbiased estimates of population parameters. For example, if test items are randomly sampled from a universe of items, the proportion of items correct in the sample of items administered to an individual is an unbiased, maximum likelihood estimate of the number of items in the universe that the individual could answer correctly (Harris, Pearlman & Wilcox, 1977). Furthermore, it does not depend on who else was tested, either as an estimate or for interpretation.

Defining universes such that random samples of items or other types of observations can be selected has great advantages but is seldom accomplished. Some have argued that it doesn't even approximate reality and that it is misleading to use a sampling model (e.g., Loevinger, 1965). At least for some limited examples, however, domains have been defined such that sampling is possible. The work of Bormuth (1968, 1970) and of Hively and his associates (e.g., Hively, 1974; Hively, Patterson & Page, 1968) provide some important examples. Some of the potential advantages of this type of approach in the area of achievement testing have been discussed by Shoemaker (1975).

Lip service is often given to the requirements of clear definition. It is fundamental to any approach that is concerned with "representative" or "random" samples of items. The idea is not new, indeed it is fairly generally accepted as desirable (see, for example, p. 28 of the APA *Standards for Educational and Psychological Tests*, 1974). It is rarely even approximated in current practice, however.

There are a lot of reasons that our measures lack the sort of rigorous definition and the sampling processes that are required. For many content areas, the problem is a very difficult one, certainly more difficult than the area of simple arithmetic problems which has become the overused exemplar. It is also the case that test publishers have relied on a totally different approach to achieving meaning. Norms rather than content definition are used to provide meaning. Note that it is not, as some have implied, the existence of norms as a basis of comparison that is the problem. Any test can be normed. Rather it is the reliance on norms as the sole basis of interpretation that is the culprit.

The need for content interpretation of achievement tests has been recognized for a long time. Ebel (1962) argued forcefully for content interpretations and provided an example where the universe and sampling procedures were defined with such rigor that it made no difference whether he or his secretary constructed the test. Ebel concluded that "a test produced by objectively defined processes may be less efficient, or lack some kinds of excellence which a creatively artistic test construction might achieve, but the increase in objective meaningfulness and reproducibility could more than offset the cost" (1962, p. 22).

Applying the principle of universe definition and sampling to measures such as observations of teacher behavior, or ratings of student art projects, is substantially more complicated than applying them to spelling or arithmetic tests. The necessity of doing so, however, is just as great if we are to achieve the sort of "objective meaningfulness and reproducibility" described by Ebel. For measures such as teacher observations the universe of admissible observations may be defined in terms of several independent characteristics, which Cronbach et al. (1972) refer to as facets. For example, one facet might be observers. A population of potential observers would need to be defined from which random samples could be selected. Other facets might be occasions, subject matter of lesson, and items on which observers make observations. In practice, it may seldom be possible to actually draw random samples of observations representing each facet. For example, location and scheduling may make some compromise necessary in the assignment of observers. Does this mean that the notions of universe definition and sampling have no place in the discussion? I think not. As noted by Cronbach:

> [I]t is customary to apply statistics derived from a random sampling model to groups of persons who were not truly drawn at random and are only loosely representative of the population of persons to which the investigation generalized. Sometimes the mismatch between model and reality is serious, sometimes not, but the problem is not peculiar to content sampling. (1971, p. 455)

Cornfield and Tukey's (1956) famous analogy between a bridge with two spans and statistical and substantive aspects of data interpretation is apropos here. The statistical span of the bridge goes from one shore to an island and the substantive span from the island to the other shore. They argue that where the substantive span is weak, it

may be better to have that span of the bridge short while stretching the statistical span.

The absence of the ability to randomly sample from a domain does not preclude the possibility of studying generalizability. Cronbach (1971) suggested a procedure that would provide a check of the content validity of a measure where components of the measure were selected judgmentally rather than by a set of sampling rules. He suggested that the content validity of such a measure could be assessed by duplicate-construction experiment. Two independent teams would use the same definition of the universe of admissible observations to construct the measures. The results of applications of the two independently constructed measures would then be compared. The ideal outcome would be that the results of the two measures were always equivalent except for sampling errors. The content validity of the measure would then be assessed in terms of the degree to which this ideal is approximated.

In the preceding comments about content validity, I have not mentioned the word competency. Is the general prescription for content validation relevant to measurement in competency-based programs? Obviously, I think it is, but some of the special characteristics of measurement in competency-based programs deserve to be considered to see if these special characteristics alter the relevance of the content validity recommendations.

Although it may be recognized that there are degrees of competency, measures are typically used to form dichotomies in competency-based programs. This may be done for purposes of certification, of licensing, or it may be done as a part of the instructional sequence with a given level of mastery required on each component, where components may or may not be arranged sequentially. The dichotomous form of the final data does not alter the need for content validity. Indeed, clear definition of the universe would seem to be an essential prerequisite for sensible judgments of "minimal standards" to be made. Additional data showing the degree of consensus on the judgments of minimal standards may be desired but these data are in addition to, rather than a substitute for, the universe definition and content validity needs. Similarly, other types of validation are needed to justify interpretations and expectations that go with the labels of competent or incompetent. As I previously argued, content validity is not sufficient but it is a necessary part of the total validation process.

Another distinguishing characteristic of competency-based programs is their emphasis on performance rather than restricting the measurement to knowledge as is often thought to happen in more traditional contexts. The emphasis is on the doing and not just the knowing. The task confronting the measurement specialist is not unlike that confronting the industrial psychologist who is attempting to construct criterion measures (see Linn, 1976). In both cases content validity is crucial. In their publication titled *Principles for Use of Personnel Selection Procedures*, the Division of Industrial-Organizational Psychology of the American Psychological Association noted that "all criteria should represent important work behaviors or work outputs" (1975, p. 3). Content validation is required to support the claim that they "represent." Clear definition of the universe of work behaviors and outputs under consideration is needed as part of that process of content validation. The universe definition is also central to making judgments regarding the claim of importance. In short, content validity is just as relevant to the measurement of performance as it is to the measurement of knowledge.

In many competency-based programs the measurement needs are so great that content validation as I've described it may be viewed as simply a pipe-dream – something that would be nice in theory but much too demanding to be of practical use in reality. It is seemingly unreasonable to expect rigorous content validation for modest sized competency-based programs, much less small ones, when such validation is lacking for even the traditional and well established tests available from the major publishers. Certainly the task is of potentially gargantuan proportions, yet it is one that has to be confronted if competency-based programs are going to even approach their promise. This is so because measurement is crucial to the conceptualization of competency-based programs.

David Krathwohl said it well in writing about performance-based teacher education (PBTE), which is considered by the committee on PBTE of the American Association of Colleges of Teacher Education to be interchangeable with competency-based teacher education (AACTE, 1974). Krathwohl argued that:

> One can predict that performance-based teacher education (PBTE) is certain to fail to meet its ultimate objective if it continues on its present course. This failure will be caused by the almost complete lack of attention given to the assessment of teaching competencies. . .(1973, p. v). Until and unless some real progress is made on resolving the problems of instrumentation and measurement,

PBTE will go down in the history books as one more bandwagon in the long line of over-simplistic solutions for complex problems (1974, p. vi).

Measurement is no less central to other competency-based programs than it is to PBTE.

Before leaving content validation, it may be worth noting that the clear universe definition is not only a requirement for content validation. I've already claimed that clear definition is also a requirement of judging the importance and appropriateness of the measure for the purposes for which it is used. Yet another function that is served by clear definition is that of publicness. An important advantage that is claimed for competency-based programs is that the participants are informed in advance of the competencies that are required and the criteria that will be used in assessing those competencies (e.g., Elam, 1971; AACTE, 1974). Such publicness requires the description of the universe of admissible observations.

Why Call It Validity?

Having described what I consider to be the two essential aspects of content validity one might question whether validity is the right designation. What I have argued for might, as Ebel (1975) suggested, better be called "content reliability." There is a close correspondence between the notions of generalizability theory and those of content validity as we have espoused them. I have retained the label of content validity for two reasons.

First, the connotation of reliability is generally narrower than what I think should be included under consideration of content validity. As commonly thought of, reliability need not involve the careful attention to unambiguous definition and sample representativeness. The common procedures for estimating various types of reliability, (e.g., internal consistency, parallel forms reliability, or test-retest reliability) can be applied with little concern for the clear definition of the limits of the domain or for questions of representativeness.

My second reason for preferring to retain the content validity label is that it makes the tie among the three types of validity more apparent. Content validity can seldom stand alone. On the other hand, it is a necessary accompaniment to the other types of validity. For example, content validity of a test may be given little consideration in a traditional criterion-related validity study but may be the central

basis for the acceptance of the criterion measure. To do otherwise could lead to the need for a criterion against which the first criterion could be validated and the continued application of that logic only leads to an infinite regress (Guion, 1977, p. 8).

Criterion-Related Validity

Criterion-referenced tests are often used to make short term instructional decisions. Commonly a dichotomy is formed and people scoring below some specified level are given some form of remediation while those above the cut-off point move on to a new set of instructional materials. As previously noted, use of scores for decisions such as this involves implicit predicitons. For example, it is assumed that those selected to move on will do better on the new material than those assigned to some form of remediation. Other implicit predictions are that students given remedial work will perform better on the test the second time around and will do better on the next segment of instruction after remediation than they would have done without it.

The implicit predictions such as the above can sometimes be supported by empirical evidence. The performance of students before and after remediation can be compared. Experiments could be conducted where some students below the cutoff score on a test are randomly selected to receive remediation while the others move on to new material. The performance of the two groups at a subsequent time could then be compared. The results of such studies would provide evidence of criterion-related validity, albeit not necessarily summarized in terms of a traditional "validity coefficient."

Hambleton and Novick (1973) suggested that the proportion of times persons above the "qualifying score" on a criterion-referenced test also are above the qualifying score on a "new test" provides an indication of validity. The new test "might well be derived from performance on the next unit of instruction, or it could be a job-related performance criterion" (p. 168).

This position is in close harmony with traditional psychometric notions of criterion-related validity. The only differences are in the metric of the scores and the type of summary statistics used. As argued by Hambleton and Novick, a dichotomous score metric is more consistent with some uses of criterion-referenced tests than is

the more finely gradated number right score and the proportion of agreement is a more useful statistic for the resulting binary variables than is a product-moment correlation.

Construct Validity

Construct validity has been a controversial topic even in the realm of psychological measurement where it was originated. Its role in educational measurement has been even more dubious. "Construct validity is not usually sought for educational tests, because they are typically thought to be valid on other grounds, namely, on the grounds of **content** validity" (Messick, 1975, p. 959). This is particularly true of competency measures.

Operational definitions are thought of as the key for competency tests and no need is seen to invoke constructs in the interpretation of performance (e.g., Harris, Pearlman & Wilcox, 1977; Osburn, 1968). But, interpretations of scores on competency tests that go beyond operational definitions are commonly made. The very word "competency" implies a construct. Even in very simple contexts the claim of competence or incompetence involves an inference. The claim may be based on a set of well defined procedures to measure the performance of an individual, but poor performance does not necessarily imply incompetence.

> The inference of inability or incompetence from the absence of correct performance requires the elimination of a number of plausible rival hypotheses dealing with motivation, attention, deafness, and so forth. Thus, a report of failure to perform would be valid, but one of inability to perform would not necessarily be valid. The very use of the term **inability** invokes constructs of attributes and process, whereas a content-valid interpretation would stick to the outcomes. (Messick, 1975, p. 960)

Inferences about competencies need to be supported by the process of construct validation. The need for construct validation is avoided only by avoiding the inferences. In some circumstances such inferences may be avoided altogether, but I think more often than not some inferences will be desired and construct validation will be called for.

Even in very simple contexts the claim of competence or incompetence involves an inference. The claim may be based on a set of well-defined procedures to measure the performance of an individual, but

poor performance does not necessarily imply incompetence. There may be alternate explanations for the poor performance such as illness or the lack of motivation (Messick, 1975). The point is simply that constructs are required to go beyond the observed performance to the inference about competency.

If the argument is accepted that construct validation is the fundamental problem in measurement for competency-based education, then, as indicated by Cronbach (1971), a variety of logical analyses and empirical studies, including experimental as well as correlational procedures, are relevant and need to be brought to bear on the problem of providing the measurement basis for inferences about incompetency. Of course, saying this hardly provides the practitioner with a clear prescription of how to proceed. Indeed, construct validation is the most demanding of the three forms of validation recognized in the *Standards for Educational and Psychological Tests* (APA, 1974). No less demanding approach is adequate for the task, however.

To try to get more specific, let me consider some of the issues that are apt to surface in a discussion of measuring competence. Discussions of competency typically emphasize performance. Knowledge is apt to be seen as a necessary component but the proof of the pudding is usually thought to be performance. Some would also add an affective component (e.g., professional attitudes) and a focus on the consequences of performance (e.g., the assessment of the teacher in terms of the performance of her/his students) (Carlson, 1974; Houston, 1974), but performance is certainly a key issue.

The advocate of competency-based education is apt to be disappointed with the garden-variety tests which may be considered useful for measuring knowledge but not for the key component of competence, i.e., performance. Unfortunately, traditional measurement is too often assumed to be synonymous with standardized, paper-and-pencil, machine-scorable, multiple-choice, norm-referenced tests.

Such a limited view of measurement is much too restrictive. The measurement of performance usually needs to involve "operant as well as repondent behavior" as recommended by McClelland (1973) for the measurement of competence. Modes other than paper-and-pencil also will often be needed. But, regardless of the mode of presentation, the appropriate basis for judging the adequacy of the measure should involve construct validity.

Personnel psychologists have had more experience with the problems of measuring performance than most educational measurement specialists. The efforts of the personnel psychologist to obtain adequate criterion measures, which, in the broad sense of the word are tests, should provide a rich source for the person about to tackle the measurement problems in a competency-based education program. The "criterion problem" in personnel selection also provides an illustration of my theme that construct validity is the fundamental issue. Developing adequate criterion measures requires detailed job or task analyses, and attention to psychometric characteristics such as generalizability, but it also requires the development of a case for the validity, generally content or construct, of the criterion measures.

During the second World War a number of well known psychologists were involved in personnel selection problems in the armed forces. One of the interesting illustrations that came from that work is a validity study designed to develop selection procedures for the Navy Gunner's Mates Schools. Surprisingly, the best predictor of grades in the Gunner's Mates Schools was a reading test. Fortunately, the response of the investigators was not simply to set up the prediction equation involving the reading test and start using it for selection. Rather, the high correlation was taken as a clue that something was wrong. A closer look at the nature of the training and the criterion measure "led to changes in the training and evaluation of students, with more emphasis being placed on assembling and disassembling guns and less on the study of manuals" (Mollenkopf, 1975). In other words, less emphasis on knowledge and more on performance.

Gulliksen and others involved in this effort for the Gunner's Mates Schools found the evidence provided by the high correlation to be inconsistent with their constructs of competence as a Gunner's Mate and of reading achievement. To have what Gulliksen (1950) later called "intrinsic validity," the relationship among the measures must make psychological sense.

Efforts to develop tests to use as criterion measures for personnel selection studies provide a number of examples that are relevant for measurement in competency-based education. Some of the best examples get pretty far removed from the stereotype of a test I mentioned above, but they do not involve different fundamental measurement issues. Mollenkopf (1975) has reviewed some of the more interesting of these efforts such as Frederiksen's "In-Basket

Test" procedure (e.g., Frederiksen, Saunders, & Wand, 1957), the work-sample tests developed in a study of possible test bias directed by Joel Campbell (see Crooks, 1972), and the situational performance tests developed by personnel psychologists at AT&T (e.g., Gael & Grant, 1972; Gael, Grant, & Ritchie, 1975a, 1975b; Grant & Bray, 1970). Tests such as these provide a desirable expansion of the view of what constitutes a test and are apt to be much closer to the type of tests that are needed in competency-based education. Ratings, checklists and a variety of other measurement techniques may also be needed. Such procedures may be viewed as nontraditional tests but the fundamental principles of measurement needed to judge the adequacy of these measures are the same that are needed for more conventional multiple-choice tests.

According to Messick (1975), "A measure estimates how much of something an individual displays or possesses. The basic question is, What is the nature of that something?" (p.957). This is the question addressed by the process of construct validation. Messick goes on to note, however, that "A second question is, Is the measured amount a lot or a little?" (p. 957). To answer this second question a standard is needed. To put it in the context of competency-based education, a standard for judging competence is needed.

As indicated by Messick, in educational measurement more attention has been devoted to the second question than the first. The two contrasting approaches to deciding whether the measured amount is a lot or a little are norm-referenced comparisons and criterion-referenced comparisons. In the former case an individual's performance is compared to that of other people. In the latter case an individual's performance is compared to a judgmentally established level or levels of performance which may be variously referred to by labels such as mastery level or minimally acceptable level.

It is not possible to provide a simple prescription for establishing construct validity. It is better thought of as a continuing process of marshalling evidence to support or refute inferences and interpretations of test results (Cronbach, 1971; Messick, 1975). The process involves logical analysis as well as a wide variety of possible empirical studies, including, for example, studies of the effects of experimental interventions.

The process of construct validation has been described most thoroughly in Cronbach's 1971 chapter in *Educational Measurement*.

He illustrated many procedures but noted that "the procedures cannot be cataloged exhaustively and no guide can tell just how to meet the requirement of hard-headed reasoning from data to conclusions" (p. 484). The emphasis, however, is on a purposeful approach starting "with a reasonably definite statement of the proposed interpretation" (p. 483). Data which can support or refute the interpretation are collected and the results are used in the refinement of the interpretation.

Instructional theory is far from the stage of providing the type of elaborated nomological net that Cronbach has described. Construct validity cannot be "established" by merely plugging a test into the network and checking on the predictions. The process is both more difficult and less clearly defined than that. On the other hand, it is only through the doing that a network can be developed and refined. In this sense the process of construct validation is just as central to the development of theory as the theory is to the validation of the measures. It is a long term iterative process.

Making progress in the validation process will require substantial effort. But, competency-based programs can ill afford to do other than invest in this effort, because, as eloquently stated by Bernard McKenna (1975), "the problem of validating competencies. . .is certainly a topic worthy of deepest consideration. . .it will remain for a long time the keystone by which the competency-based movement stands or becomes a house of cards." (pp. 38-39)

Chapter VI

Messages For Educators

Much of the current writing regarding the effects of competency-based programs has been centered on the students. However, the burden of change is also on the practice and theory of educational measurement. School people have a long history of using testing results as supportive information in decision making concerning individual students. The new programs place the educator in the position of making a decision concerning a student based on a single result. This position is uncomfortable, given our history.

The discomfort concerning the use of tests is further complicated by a series of terms which have unclear meanings. The lack of clarity in the use of words may be due to a lack of empirical investigation of the problem. Clarity will not necessarily come with time, nor will time alone bring empirical investigation. The papers by Hecht and Pottinger are witness to the fact that time has not brought clarity to the fields which have a long history in certification and licensing. Both of those papers use the term "competence" in the sense of a construct – in fact, competency is a highly complex behavioral construct. The concentration on the construct definition and the complexity of behavior gives an impo tant message to educators.

If educators concentrate on the alternative definition, i.e., the cut-off score, the complexity of the behavior measured will never be addressed. Using the standard of performance definition allows one to consider the level of behavior which is acceptable, but not whether the behavior is the acceptable behavior.

The problem of the cut-off score or the standard of performance definition of the word competency is shown in the Shepard, Jaeger, and Conaway papers. Each shows how the act of setting a standard is currently a judgmental act, but they differ with respect to the strength and level of inference one would make on the basis of the competency-based measure. The level and kind of inference one wants to make on the basis of the measure is also the theme that runs through both of the Linn papers concerning the technical quality of measurement instruments.

Linn also shows how confusion about language is not unique to "competency" in that the sources for his review are drawn from documents which concern objective-referenced instruments, criterion-referenced instruments, mastery tests, and domain-referenced tests. In the last fifteen years new descriptors of tests have grown up which by their title seem to be outside the realm of classical test theory. Proponents of these various forms of testing often bemoan the fact that classical test theory evolved out of individual difference psychology and hold their candidate for replacement as a better conception of human behavior. However, as Shepard points out, the alternatives also are based on a psychology of individual learning, i.e., linear and hierarchical. Also, Linn shows how techniques of classical test theory can be applied to these alternatives.

A message for educators at all levels, then, is to proceed very carefully. Be sure that the philosophy and psychology implicit in a competency-based program is consistent with the values of the educational institution before you decide to implement the program. But suppose that the choice is not whether or not to implement the program. The choice in some states and cities is not, " Should we or should we not?" but, "How should we implement?" If this is the case (and as noted earlier in 1978 this was the case in 33 states), then what can be done? The message of this document is to proceed along three different routes.

First, the "competencies" in terms of a construct must be defended. The defense can take many forms, but clearly the "what" of what is taught in the schools must be defensible. This task belongs to all educators, but most centrally it is the concern of curriculum specialists and administrators. Implementation of a competency-based program should not imply that the goals of the school will be determined by a test construction expert.

The second step is to find an instrument which measures the construct in a defensible manner. Whether the traditional indices of reliability and validity are used to judge the quality of information produced by the instrument is of interest here, but not of central concern. The concern for which indices are used is clearly central to the psychometrician. However, school testing directors and directors of state assessment programs must also worry the question of how good an instrument is, without the benefit of ongoing research in the

area. Psychometricians have a responsibility to communicate with people who are forced to make decisions concerning instruments. But also, the users of tests must see that settling for no information concerning an instrument (because the classical indices have been called into question) is not defensible behavior.

The third step is to decide upon the standard of performance required of the students. The procedures used in this step may be those recommended by Jaeger, Shepard or Conaway. However, it is the final step and not the first. All educators have a responsibility to show the public and the legislature the folly of setting standards before you know what you are measuring and how well you are measuring it.

References

Agree, B.C. The threat of institutional licensure. *American Journal of Nursing*, 1973, *73*, 1758-1763.

Airasian, P.W., & Madaus, G.F. Criterion-referenced testing in the classroom. *NCME Measurement in Education*, 1972, *3* (4), 1-8.

Albemarle Paper Co. et al. v. *Moody et al.*, 45 L Ed 2nd 280, 95 S Ct 2362 (June 1975).

American Association of Colleges for Teacher Education, Committee on Performance Based Teacher Education. *Achieving the potential of performance based teacher education: Recommendations*. Washington, D.C.: Author, 1974.

American Nurses Association. *Certification examination, psychiatric-mental health nursing — Bulletin of information for candidates 1974*. Princeton, N.J.: Educational Testing Service, 1974.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. *Standards for educational and psychological tests*. Washington, D.C.: American Psychological Association, Inc., 1974.

American Psychological Association, Division of Industrial-Organizational Psychology. *Principles for the validation and use of personnel selection procedures.* Dayton, Ohio: The Industrial-Organizational Psychologist, 1975.

Anastasi, A. *Levels of competence—minimal or optimal.* Presentation made at the National Conference on Evaluating Competence in the Health Professions, sponsored by the Professional Examination Service, New York, November, 1976.

Andrews, B.J., & Hecht, J.T. A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement*, 1976, *36*, 45-50.

Angoff, W. Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Block, J.H. (Ed.), *Mastery learning: Theory and practice.* New York: Holt, Rinehart and Winston, 1971.

Bloom, B. Mastery learning. In J.H. Block (Ed.), *Mastery learning: Theory and practice.* New York: Holt, Rinehart and Winston, 1971. (a)

Bloom, B. Mastery learning and its implications for curriculum development. In E.W. Eisner (Ed.), *Confronting curriculum reform.* Boston: Little, Brown, 1971. (b)

Bormuth, J. R. Cloze test reliability: Criterion reference scores. *Journal of Educational Measurement*, 1968, *5*, 189-196.

Bormuth, J.R. *On the theory of achievement test items.* Chicago: University of Chicago Press, 1970.

Boyd, J.L., Jr. *Development of the testing program for the Academy of Certified Social Workers.* Princeton, N.J.: Educational Testing Service, 1975.

Bray, D.W. The assessment center method of appraising management potential. In J.W. Blood (Ed.), *The personnel job in a changing world.* New York: American Management Association, 1964.

Bray, D. W., & Campbell, R.J. Selection of salesmen by means of an assessment center. *Journal of Applied Psychology*, 1968, *52*, 36-41.

Bray, D.W., & Grant, D.L. The assessment center in the measurement of potential for business management. *Psychological Monographs,* 1966, *80* (17, Whole No. 625).

Brennan, R.L., & Kane, M.T. An index of dependability for mastery tests. *Journal of Educational Measurement*, 1977, *14*, 277-289.

Brickell, H. *The influence of external political factors on the role and methodology of evaluation.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1974.

Byham, W.C. Assessment centers for spotting future managers. *Harvard Business Review*, 1970, *48* (4), 150-160.

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, *56*, 81-105.

Campbell, J., Dunnette, M., Lawler, E., & Weick, K. *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill, 1970.

Carlson, A.B. *Criteria for competence in the legal profession.* Presentation made at the conference on the Future of Law School Admission Council Research, San Francisco, September 1974.

Carlson, A.B., & Werts, C.E. *Relationships among law school predictors, law school performance and bar examination results.* Unpublished manuscript, Educational Testing Service, 1975.

Carroll, J. Problems of measurement related to the concept of learning for mastery. In J.H. Block (Ed.), *Mastery learning: Theory and practice.* New York: Holt, Rinehart and Winston, 1971.

Carver, R.P. Special problems in measuring change with psychometric devices. In *Evaluative research: Strategies and methods.* Pittsburgh, Penn.: American Institutes for Research, 1970.

Carver, R.P. Two dimensions of tests: Psychometric and edumetric. *American Psychologist*, 1974, *29*, 512-518.

Clark, J.P., & Thomson, S.D. *Competency tests and graduation requirements.* Reston, Va.: National Association of Secondary School Principals, 1976.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, *20*, 37-46.

Cohen, J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement of partial credit. *Psychological Bulletin*, 1968, *70*, 213-220.

Cornfield, J., & Tukey, J.W. Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 1956, *27*, 907-949.

Cox, R.C., & Graham, G.T. The development of a sequentially scaled achievement test. *Journal of Educational Measurement,* 1966, *3,* 147-150.

Cronbach, L.J. *Essentials of psychological testing* (3rd ed.). New York: Harper and Row, 1970.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Cronbach, L.J. Dissent from Carver. *American Psychologist,* 1975, *30,* 602-603.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley, 1972.

Crooks, L.A. (Ed.). *An investigation of sources of bias in the prediction of job performance — A six year study.* Princeton, N.J.: Educational Testing Service, 1972.

Dodd, W.E. Will management assessment centers insure selection of the same old types? *Proceedings of the 78th Annual Convention of the American Psychological Association,* 1970, *5,* 569-570. (Summary)

Dowaliby, F.J., & Andrew, J. *Relationships between clinical competence ratings and examination performance.* Philadelphia, Penn.: National Board of Medical Examiners, n.d.

Dunnette, M.D. Multiple assessment procedures in identifying and developing managerial talent. In P. McReynolds (Ed.), *Advances in psychological assessment* (vol.2). Palo Alto, Calif.: Science and Behavior Books, 1971.

Ebel, R.L. Content standard test scores. *Educational and Psychological Measurement,* 1962, *22,* 15-25.

Ebel, R.L. The value of internal consistency in classroom examinations. *Journal of Educational Measurement,* 1968, *5,* 71-73.

Ebel, R.L. *Essentials of educational measurement*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Ebel, R.L. *Measuring educational achievement*. Englewocd Cliffs, N.J.: Prentice-Hall, 1974.

Ebel, R.L. *Prediction? Validation? Construct validity?* Paper presented at Content Validity II, a conference at Bowling Green State University, July 1975.

Elam, S. *Performance-based teacher education: What is the state of the art?* Washington, D.C.: American Association of Colleges for Teacher Education, 1971.

Everitt, B.S. Moments of the statistics Kappa and weighted Kappa. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 97-103.

Fine, S.A., & Wiley, W. *An introduction to functional job analysis.* Kalamazoo, Mich.: Upjohn Institute, 1971.

Fitzpatrick, R., & Morrison, E.J. Performance and product evaluation. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Fleiss, J.L., & Cohen, J. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 1973, *33*, 613-619.

Fleiss, J.L., Cohen, J., & Everitt, B.S. Large sample standard errors of Kappa and weighted Kappa. *Psychological Bulletin*, 1969, *72*, 323-327.

Fortune, J.C. *Professional licensing and certification: Analysis of testing procedures and their implications for practice*. Presentation made at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Franklin, J.L. *A path analytic approach to describing causal relationships among social-psychological variables in multilevel organizations*. Ann Arbor, Mich.: Institute for Social Research, 1973.

Frederiksen, N., Saunders, D.R., & Wand, B. The in-basket test. *Psychological Monographs*, 1957, *71* (9, Whole No. 438).

Gael, S., & Grant, D.L. Employment test validation for minority and non-minority telephone company service representatives. *Journal of Applied Psychology*, 1972, *56*, 135-139.

Gael, S., Grant, D.L., & Ritchie, R.J. Employment test validation for minority and non-minority telephone operators. *Journal of Applied Psychology*, 1975, *60*, 411-419. (a)

Gael, S., Grant, D.L. & Ritchie, R.J. Employment test validation for minority and non-minority clerks with work sample criteria. *Journal of Applied Psychology*, 1975, *60*, 420-426. (b)

Gagné, R.M. *The conditions of learning*. New York: Holt, Rinehart and Winston, 1965.

Gagné, R.M., & Paradise, N.E. Abilities and learning sets in knowledge acquisition. *Psychological Monographs*, 1961, *75*, (14, Whole No. 518).

Ghiselli, E.E. Dimensional problems of criteria. *Journal of Applied Psychology*, 1956, *40*, 1-4.

Ghiselli, E.E. *The validity of occupational aptitude tests.* New York: Wiley, 1965.

Galser, R. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist,* 1963, *18*, 519-521.

Glass, G.V. *Standards and criteria*, (Occasional Paper Series, No. 10). Kalamazoo, Mich.: Western Michigan University, Evaluation Center, 1977.

Grant, D.L, & Bray, D.W. Validation of employment tests for telephone company installation and repair occupations. *Journal of Applied Psychology*, 1970, *54*, 7-14.

Guion, R.M. Content validity – The source of my discontent. *Applied Psychological Measurement*, 1977, *1*, 1-10.

Gilliksen, H. Intrinsic validity. *American Psychologist*, 1950, *5*, 511-517.

Haladyna, T.M. On the psychometric-edumetric dimensions of tests. *American Psychologist*, 1975, *30*, 603-604.

Hambleton, R.K., & Novick, M.R. Toward an integration of theory and method for criterion referenced tests. *Journal of Educational Measurement*, 1973, *10*, 159-170.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D.B. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, *48*, 1-47.

Harris, C.W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. *Journal of Educational Measurement*, 1972, *9*, 27-29.

Harris, C.W. Some technical characteristics of mastery tests. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974.

Harris, C.W., & Pearlman, A.P. Conventional significance tests and indices of agreement or association. In C.W. Harris, A.P. Pearlman & R.R. Wilcox (Eds.), *Achievement test items — Methods of Study* (CSE Monograph Series in Evaluation, No. 6). Los Angeles: University of California, Center for the Study of Evalu- of Evaluation, 1977.

Harris, C.W., Pearlman, A.P. & Wilcox, R.R. (Eds.), *Achievement test items — Methods of study* (CSE monograph Series in Evaluation, No. 6). Los Angeles: University of California, 1977.

Heath, D.H. What the enduring effects of higher education tell us about a liberal education. *Journal of Higher Education*, 1976, *47*, 173-190.

Hecht, K.A. *Validation of professional licensing and certification examinations: A methodological dilemma*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, 1974. (ERIC Document Reproduction Service No. ED 110 481)

Hecht, K.A. *Current status and methodological problems of validating professional licensing and certification exams*. Presentation made at the Annual Meeting of the National Council of Measurement in Education, San Francisco, April 1976.

Hemphill, J.K. *A summary of the engineering study*, Princeton, N.J.: Educational Testing Service, 1963.

Hively, W., Maxwell, G., Rabehl, G., Stension, D., & Lundin, S. *Domain-referenced curriculum evaluation: A technical handbook on a case study from the MINNEMAST project* (CSE Monograph Series in Evaluation, No. 1) Los Angeles: University of California, Center for the Study of Evaluation, 1973.

Hively, W. (Ed.) *Domain-referenced testing*. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Hively, W., Patterson, H., & Page, S. A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, *5*, 275-290.

Hoffman, P.J. *Continued competency assurance: Some research and measurement considerations*. Presentation made at the National Conference on Evaluating Competence in the Health Professions, sponsored by the Professional Examination Service, New York, November 1976.

Hogan, D.B. *State licensing, academic credentials and accreditation: Implications and recommendations for postsecondary educational institutions*. Presentation made at the Conference on Credentialism, University of California Law School, Berkeley, April 1977.

Horn, J.L. Some characteristics of classroom examinations. *Journal of Educational Measurement*, 1966, *3*, 293-295.

Horn, J.L. Is it reasonable for assessments to have different psychometric properties than predictors? *Journal of Educational Measurement*, 1968, *5*, 75-77.

House, E., Rivers, W., & Stufflebeam, D. *An assessment of the Michigan accountability system* (Evaluation Report Series, No. 2). Kalamazoo, Mich.: Western Michigan University, Evaluation Center, 1975.

Houston, W.R. Competency based education. In W.R. Houston (Ed.), *Exploring Competency based education*. Berkeley, Calif.: McCutchan, 1974.

Hoyt, D.P. *The relationship between college grades and adult achievement, a review of the literature* (Research Report No. 7). Iowa City, Iowa: American College Testing Program, 1965.

Hubbard, J.P. Programmed testing in the examinations of the National Board of Medical Examiners. In 1963 *Invitational conference on testing problems*. Princeton, N.J.: Educational Testing Service, 1964.

Hubbard, J.P., Levit, E.J., Schumacher, C.F., & Schnabel, T.G. An objective evaluation of clinical competence. *The New England Journal of Medicine*, 1965, *272*, 1321-1328.

Huck, J.R. Assessment centers: A review of external and internal validities. *Personnel Psychology*, 1973, *26*, 191-212.

Huff, S. Education, work and competence. *Society*, 1976, *13* (2), 44-51.

Huynh, H. On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 1976, *13*, 253-264.

Huynh, H. *The Kappamax reliability index for decisions in domain-referenced testing*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1977.

Jensen, O. *Some thoughts on minimum competency interpretations of performance on (M-C Item) achievement tests*. Princeton, N.J.: Educational Testing Service. 1972.

Jaeger, R. *Measurement consequences of selected standard-setting models*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Jones, J. ETS considers testing and the public interest. *American Psychological Association Monitor*, 1977, *8*, 4.

Klein, S.B., & Kosecoff, J. *Issues and procedures in the development of criterion-referenced tests* (ERIC Clearinghouse on Tests, Measurement, and Evaluation, TM report 26). Princeton, N.J.: Educational Testing Service, 1973.

Klemp, G.O. *Three factors of success in the world of work: Implications for curriculum in higher education*. Presentation made at the Annual Convention of the American Association of Higher Education, Chicago, March 1977.

Kosecoff, J., & Fink, A. The feasibility of using criterion-referenced tests. In *Measure of student growth for the sustaining effects study*. Santa Monica, Calif.: Systems Development Corporation, 1976.

Krathwohl, D. Introductory note. In J.C. Merwin, *Performance-based teacher education: Some measurement and decision-making considerations*. Washington, D.C.: Association of Colleges of Teacher Education, 1973.

Ledvinka, J. & Schoenfeldt, L.F. Legal developments in employment testing: *Albemarle* and beyond. *Personnel Psychology*, 1978, *31*, 1-13.

Light, R.J. Issues in the analysis of qualitative data. In R.M.W. Travers (Ed.), *Second handbook of research on teaching*. Chicago: Rand McNally, 1973.

Likert, R., & Bowers, D.G. Organizational theory and human resource accounting. *American Psychologist*. 1969, *24*, 585-592.

Likert, R., & Bowers, D.G. Improving the accuracy of P/L reports by estimating the change in dollar value of the human organization. *Michigan Business Review*, 1973, *25* (2), 15-24.

Lindquist, E.F. The theory of test construction. In H.E. Hawkes, E.F. Lindquist, & C.R. Mann (Eds.), *The construction and use of achievement examinations*. Boston: Houghton-Mifflin, 1936.

Linn, R.L. *Construct validity and measurement of competency-based education*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1976.

Livingston, S.A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, *9*, 13-26. (a)

Livingston, S.A. A reply to Harris' "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." *Journal of Educational Measurement*, 1972, *9*, 31. (b)

Livingston, S.A. Reply to Shavelson, Block and Ravitch's "Criterion-referenced testing: Comments on reliability." *Journal of Educational Measurement*, 1972, *9*, 139-140. (c)

Livingston, S.A. *The criterion-referenced reliability of a single measure* (COPA-76-01). Princeton, N.J.: Educational Testing Service, 1976.

Loevinger, J. Person and population as psychometric concepts. *Psychological Review*, 1965, *72*, 143-155.

Lord, F.M., & Novick, M.R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Loveland, E.H. *Alternatives and innovations*. Presentation made at the National Conference on Evaluating Competence in the Health Professions, sponsored by the Professional Examination Service, New York, November 1976.

MacKinnon D.W. *An overview of assessment centers* (Technical Report No. 1). Greensboro, N.C.: Center for Creative Leadership, 1975.

Marlowe, J. Testing, testing . . . can you hear me? *Phi Delta Kappan*, 1976, *58*, 256-257.

Marshall, J.L. *The mean split-half coefficient of agreement and its relation to other test indices: A study based on simulated data* (Technical Report 350). Madison: University of Wisconsin, Research and Development Center for Cognitive Learning, 1975.

Marshall, J.L., & Haertel, E.H. *A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement*. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C., 1975.

Maslow, A.P. Licensing tests – Occupational bridge or barrier? *Proceedings third annual conference*, Council on Occupation Licensing, July 1971. (see Shimberg, 1972)

Maxwell, G., Hively, W., Lundin, S., Rabehl, G., & Sension, D. *Curriculum evaluation in the MINNEMAST project: A case study in domain-referenced testing.* Minneapolis: University of Minnesota, 1971.

McClelland, D.C. Testing for competence rather than for "intelligence." *American Psychologist*, 1973, *28*, 1-14.

McClelland, D.C. *Power: The inner experience*. New York: Irvington Press, 1975.

McClelland, D.C. Personal communication, 1976.

McClelland, D.C., & Burnham, D.H. Power in the great motivator. *Harvard Business Review*, 1976, *54* (2), 100-110.

McClelland, D.C. & Dailey, D.H. *Professional competencies of human service workers: A report to the College for Human Services*. Boston, Mass.: McBer and Company, 1974.

McGuire, C.H. A process approach to the construction and analysis of medical examinations. *Journal of Medical Education,*, 1963, *38*, 556-563.

McGuire, C.H. A scientific approach to problems of professional assessment. *The Canadian Medical Association Journal*, 1969, *100*, 593-598.

McGuire, C.H., & Babbot, D. Simulation techniques in the rneasurement of problem-solving skills. *Journal of Educational Measurement*, 1967, *4*, 1-10.

McKenna, B.H. An essay on "politics of competence." In J.G.G. Merrow II (Ed.), *The politics of competence: A review of competency-based teacher education*. Washington, D.C.: National Institute of Education, 1975.

Merwin, J.C. *Performance-based teacher education: Some measurement and decision making considerations*. Washington, D.C.: American Association of Colleges of Teacher Education, 1973.

Meskauskas, J.A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research*, 1976, *45*, 133-158.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, *30*, 955-966.

Michigan Department of Education. *Interpretive manual: 1976-1977*. Lansing: Author, 1976. (a)

Michigan Department of Education. *Status report on Michigan basic skills education: 1976-1977*. Lansing: Author, 1976. (b)

Millman, J. *Tables for determining the number of items needed on domain-references tests and the number of students to be tested* (Technical Paper No. 5). Los Angeles: Instructional Objectives Exchange, April 1972.

Millman, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 1973, *43*, 205-216.

Millman, J. Criterion-referenced measurement. In W.J. Popham (Ed.), *Evaluation in education*. Berkeley, Calif.: McCutchan, 1974.

Millman, J., & Popham, W.J. The issue of item and test variance for criterion-referenced tests: A clarification. *Journal of Educational Measurement*, 1974, *11*, 137-138.

Mollenkopf, W.G. *Some observations on validity and validation*. Presidential Address to the Division of Evaluation and Measurement of the American Psychological Association at the Annual Meeting, Chicago, August 1975.

Moody et al. v. *Albemarle Paper Co. et al.*, 474 F. 2d 134 (4th Cir. 1973).

Moyer, J.E. *A comparison of the Kuder-Richardson formula and Kappa as estimates of the reliability of criterion-referenced tests* (Doctoral dissertation, Western Michigan University, 1976). *(Dissertation Abstracts International*, 1976, *37*, 2962-B. (University microfilms No.76-26, 756)

Moyer, J.E. *Alternative reliability indices*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, 1977.

National Association of Social Workers, Inc. *Academy of Certified Social Workers—Information bulletin 1975*. Princeton, N.J.: Author and Educational Testing Service, 1975.

National Board of Medical Examiners. *National examination program for assistants to the primary care physician—Project report 1973-1974*. Philadelphia, Penn.: National Board of Medical Examiners, n.d.

National Education Association. *Guidelines and cautions for considering criterion-referenced tests*. Washington, D.C.: Author, 1975.

Nedelsky, L. Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 1954, *14*, 3-19.

Noe, M.J., & Algina, J. *An investigation of a single administration estimate of a criterion-referenced reliability index*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1977.

Novick, M.R., & Jackson, P.H. Bayesian guidance technology. *Review of Educational Research*, 1970, *40*, 459-494.

Novick, M.R., & Jackson, P.H. *Statistical methods for educational and psychological research*. New York: McGraw-Hill 1974.

Novick, M.R., Lewis, C., & Jackson, P.H. The estimation of proportions in m groups. *Psychometrika*, 1973, *38*, 19-46.

Osburn, H.G. Item sampling for achievement testing. *Educational and Psychological Measurement*, 1968, *28*, 95-104.

Popham, W.J., & Husek, T.R. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, *6*, 1-9.

Pottinger, P. Techniques and criteria for designing and selecting instruments for assessing teachers. In B. Levitov (Ed.), *Licensing and accreditation in education: The law and the state interest*. Lincoln, Neb.: Study Commission on Undergraduate Education and the Education of Teachers, 1976.

Price, et al. *Measurement and predictors of physician performance*. Salt Lake City, Utah: Aaron Press, 1971.

Quirk, T.J. Some measurement issues in competency-based teacher education. *Phi Delta Kappan*, 1974, *55*, 316-319.

Quirk, T.J., Witten, B.J., & Weinberg, S.F. *A critical review of research related to the National Teacher Examinations*. Princeton, N.J.: Educational Testing Service, 1972.

Reader's guide to periodical literature, 1973-1974, *33*.

Ryans D.G., & Frederiksen, N. Performance tests of educational achievement. in E.F. Lindquist (Ed.), *Educational measurement*. Washington, D.C.: American Council of Education, 1951.

Schumacher, C.F. Validation of the American Board of Internal Medicine Written Examination. *Annals of Internal Medicine*, 1973, *58*, 131-135.

Sharon, A.T. *Assessment of competence in social work: A report on the 1973 assessment results of the Academy of Certified Social Workers (ACSW) program*. Princeton, N.J.: Educational Testing Service, 1974.

Shavelson, R., Block, J., & Ravitch, M. Criterion-referenced testing: Comments on reliability. *Journal of Educational Measurement*, 1972, *9*, 133-137.

Shimberg, B. *Certification and licensing in competency-based education: the challenge for measurement*. Presentation made at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Shimberg, B., Esser, B.F., & Kruger, D.H. *Occupational licensing and public policy*. Princeton, N.J.: Educational Testing Service, 1972.

Shoemaker, D.M. Note on the attenuating effect of zero-variance items on KR-20. *Journal of Educational Measurement*, 1969, *6*, 255-256.

Shoemaker, D.M. Toward a framework for achievement testing. *Review of Educational Research*, 1975, *45*, 127-147.

Simon, G.B. Comments on "Implications of criterion-referenced tests." *Journal of Educational Measurement*, 1969, *6*, 259-260.

Smilansky, J., & Guerin, R.O. *Minimal acceptable performance levels for criterion-referenced multiple choice examinations and their validation*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1976.

Spady, W.G. Competency-based education: A bandwagon in search of a definition. *Educational Researcher*, 1977, *6* (1), 9-14.

Spady, W.G., & Mitchell, D.E. Competency-based education: Organizational issues and implications. *Educational Researcher*, 1977, *6* (2), 9-15.

Stanley, J.C. Reliability. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Strasler, G.M., & Raeth, P.G. *An internal consistency estimate for criterion-referenced tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, 1977.

Stuit, D.B. (Ed.). *Personnel research and test development in the U.S. Bureau of Naval Personnel* (Project N-106 and the College Entrance Examination Board). Princeton, N.J.: Princeton University Press, 1947.

Subkoviak, M. Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 1976, *13*, 265-276.

Swaminathan, H., Hambleton, R.K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 1974, *11*, 262-267.

Torgerson, W.S. *Theory and methods of scaling*. New York: Wiley, 1958.

U.S. Government. Guidelines for employee selection procedures (Title 29 Labor, Chap. 14, Equal Employment Opportunity Commission). *Federal Register*, July 1975, Pt. 1607, pp. 807-812.

Weiss, I.R., & Conaway, L.E. The establishment of student performance criterion levels for judging the acceptability of assessment results. In *Formal papers and selected bibliography* (report of the Sixth Annual Conference on Large Scale Assessment). Denver, Colo.: National Assessment of Educational Progress, 1976.

White, D.M. Title VII and the masters of reality: Eliminating credentialism in the American labor market. *Georgetown Law Journal*, 1976, *64*, 1213-1244.

White, W. Testing and equal opportunity. *The Civil Rights Digest*, 1975, *7* (3), 42-51.

Woodson, M.I.C.E. The issue of item and test variance for criterion-referenced tests. *Journal of Educational Measurement*, 1974, *11*, 63-64. (a)

Woodson, M.I.C.E. The issue of item and test variance for criterion-referenced tests: A reply. *Journal of Educational Measurement*, 1974, *11*, 139-140. (b)

J.E.M. Special Issue

NCME announces the publication of another special issue of the *Journal of Educational Measurement*. This issue of "Setting Standards" was planned and supervised by JEM editor, Lorrie Shepard, with the assistance of contributors and reviewers. Articles included in the issue are:

Societal Standards Nancy Burton
Standards and Criteria Gene Glass
Demands, Cautions, and Suggestions for Setting Standards Robert L. Linn
Educational Performance Standards: Image or Substance Henry M. Levin
Responses by James H. Block, Ronald K. Hambleton, W. James Popham, and Michael Scriven

Make sure of the copies you need by writing to: J.E.M. Special Issue, 1230 17th Street, N.W., Washington D.C. 20036

J.E.M., V. 15, No. 4 "Setting Standards"

Prepaid orders*

Single copy @ $5.00
2-49 copies @ $4.75
50-99 copies @ $4.50
100 copies @ $4.00

* Rates are for NCME members **only**. All others add $2.00 to the price of each copy. **Orders must be prepaid.**